TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

KHAZAR KHORRAMI

CANCER DETECTION FROM HISTOPATHOLOGY IMAGES

Master of Science thesis

Examiner: Assoc. Prof. Heikki Huttunen
Examiner and topic approved by the Faculty Council of the Faculty of Computing and Electrical Engineering on 29th March 2017

ABSTRACT

KHAZAR KHORRAMI: Cancer Detection from Histopathology Images
Tampere University of Technology
Master of Science thesis, 51 pages
May 2018
Master's Degree Programme in Information Technology
Major: Signal Processing
Examiner: Assoc. Prof. Heikki Huttunen
Keywords: Lung cancer, computer-aided diagnosis, machine learning, histopathology image analysis

Histopathological analysis of whole-slide images is one of the most widely used techniques for diagnosis of lung cancers. In this study, a fully automated pipeline was developed to detect cancer from histopathology slides of lung tissue. We obtained 1067 histopathology images of lung adenocarcinoma and 1060 images of squamous cell carcinoma from the legacy archive of The Cancer Genome Atlas (TCGA) dataset and used them to test the proposed methodology. At preprocessing step, we trained a classification model to detect clinically relevant patches of images using statistical measurements. In the next step, cells and nuclei of the cells were segmented and various texture and morphology features were extracted from images and segmented objects. At the final step, different classification models were applied to distinguish between malignant tissues and adjacent normal cells. The results indicates that the usage of machine learning algorithms at pre-processing step for detecting relevant sections of whole slide images improves the performance of automated cancer detection systems substantially.

PREFACE

I would first like to extend my deepest thanks to my thesis advisor and examiner Associate Prof. Heikki Huttunen from Signal Processing department at Tampere University of Technology for all of his technical and financial supports during this work. He was kindly and openly helping me whenever I faced any trouble or question during research or writing process. He provided me the opportunity to develop my own ideas on the topic but guided me through the right direction whenever needed.

I would also like to thank Pekka Ruusuvuori from Prostate Cancer Research Center at University of Tampere who helped me on the biological concepts of the work. We had several meetings and I received great help and comments from him at various steps of the work from data acquisition and annotation to designing the appropriate feature extraction pipeline in CellProfiler software.

Finally, I would like to express my great thanks to my father, Mohammad and my uncle, Darioush and my wonderful friends Alireza Zare, Razieh Zare, Zeinab Rezaei, Zahra Abbaszadeh, Nahid Sheikhi pour, Saboktakin Hayati, Vida Arabzadeh, Ali Zare, Alireza and Davood Rasti, Deepa Naik, Lingyu Zhu, and many others who supported me during my master's studies at Tampere University of Technology.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND SYMBOLS

WSI	Whole slide imaging
CAD	Computer-aided diagnosis
ML	Machine learning
HCA	High-content analysis
HE	Hematoxylin and eosin
SVM	Support vector machine
TMA	Tissue micro-array
LBP	Local binary patterns
LPQ	Local phase quantization
DNA	Deoxyribonucleic acid
GLCM	Gray-level co-occurrence matrix
LDA	Linear discriminant analysis
DT	Decision tree
RF	Random forest
ROC	Receiver operator characteristic
AUC	Area under the curve
CV	Cross validation
TCGA	The Cancer Genome Atlas
GDC	Genomic Data Commons
NCI	National Cancer Institute
DL	Deep learning
CNN	Convolutional neural network
GPU	Graphics processing unit

1. INTRODUCTION

1.1 Motivation

Cancer causes thousands of deaths worldwide each year. However, early detection and accurate diagnosis can improve the survival rate of suffering patients significantly. Histopathological analysis of microscopic whole slide images (WSIs) is one of the most widely used techniques for diagnosis of cancer. However, qualitative examination of huge amount of data presented in whole-slide images is a laborious task demanding hours of professional work. Moreover, visual inspection of images by experts can lead to inaccurate results due to its subjective nature. Computer-aided diagnosis (CAD), on the other hand, enables fast and accurate analysis of huge amount of data presented in microscopic WSIs and thus, provides significant advantages over traditional methods [1][2].

1.2 Automated image cytometry

Applied in the diagnosis of some diseases such as cancer, image cytometry is the technique of investigating microscopic images for their biologically interpretable features. In this method, different visual features of cells and nuclei of the cells are measured from microscopic images of tissue samples. These measurements help pathologists to assign different degrees of abnormality to examine tissues and categorize them into various pathological classes. Measured features include attributes that undergo variations in presence of the suspected disease, such as size and morphology of the cells and the nuclei of cells. In the most image cytometry techniques, images obtained through optical microscopy are stained prior to further analysis to enhance their contrast and visualize the different image objects and components.

Automated image cytometry techniques range from simple cell counting to more sophisticate methods such as High-content analysis (HCA) applied for detecting smaller substances like peptides. These quantitative techniques have considerable advantages over traditional subjective methods; Firstly, human observation points a few qualitative features while image cytometry applies different measurements

to detect numerous quantitative attributes such as size, shape, and count of cells and sub-cellular components. Some delicate features such as a 10 percent change in nucleus size can be revealed only through exact measurements. Furthermore, image cytometry can study huge amount of data such as thousands of images in a reasonable time whereas subjective analysis of numerous microscopic images is impossible or very time-consuming task.[10]

1.3 Related works

A typical computerized model of automated diagnosis systems consists of three main phases including preprocessing of WSIs, measuring various image features, and applying Machine learning (ML) algorithms to classify different malignancy patterns.

The big visual data presented in the hitopathology slides imposes a lot of redundant work for experts as well as CAD systems. A number of researches have aimed to develop a method for distinguishing diagnostically relevant regions within pathology WSIs [3][4] based on various numerical features of the images such as color, texture, and shape descriptors. The results of such studies help to provide a faster work-flow for cancer diagnosis systems through recognizing and emphasizing several important regions of interest within histopathology slides.

Furthermore, enhancing the quality of images obtained from whole slide scanners is an important section of any automated system designed for analysis of such images. Various image processing techniques such as color enhancement methods can be applied to increase the contrast of images and visualize objects of interest such as cells and nuclei of the cells. In order to increase the image contrast and make tissue components more visible, histopathology slides are stained using different methods. In hematoxylin and eosin (HE) staining method, nuclei of the cell appear in blue (or purple) color and cytoplasm appear in red (or pink).

Bahlman et al. [4] firstly separated hematoxylin and eosin channels to segment tissue components (nuclei and cytoplasm). Next, using HE color components, they calculated the statistical distribution of nuclei and cells within smaller patch of size 256 x 256 pixels. Finally, they applied a linear classification method to recognize between clinically relevant and irrelevant regions the whole slide histopathology images.

Peikari et al. [3] applied a texture-based analysis method for triaging diagnostically important regions of pathology WSIs. They measured statistical features from ran-

domly selected image patches and used a support vector machine (SVM) classifier to distinguish between clinically relevant and irrelevant patches.

Moreover, detecting and removing image artifacts is an important task of the preprocessing step. One of the main common image artifacts presented in the histopathology images are the regions in which tissue is folded twice or more while placing on the microscope slide. Some studies have employed color enhancement techniques to automatically detect these regions within WSIs [6][5].

In general, increasing the saturation component of an image helps to improve the image quality and emphasize objects of interest. Moreover, tissue folds are distinguishable with their high saturation and low intensity. In order to locate tissue fold areas, Palokangas et al. [5] subtracted intensity channel from saturation channel to obtain an image in which tissue folds are remained and other normal sections are disappeared. In another work, Butista and Yagi [6] introduced an adaptive shifting metric using the difference between luminance and saturation channels to emphasize and locate pixels of tissue folds areas within low-pixel resolution version of WSIs.

Moreover, a variety of recent studies aimed to develop an automated pipeline to detect cancerous tumors using automated analysis of histopathology WSIs [7][8][9]. Ojansivu et al. [7] tried to classify morphological patterns of breast cancer in a data set of whole-slide tissue micro-array (TMA) samples. They employed local binary patterns (LBP) and local phase quantization (LPQ) texture descriptors and SVM classifier to categorize the TMA samples into three classless of extensive tubular formation, intermediate tubular formation, and no tubular formation.

Kumar et al. [8] proposed a pipeline for automated detection of cancer from microscopic biopsy images. In the first step, they enhanced the images using histogram equalization methods and segmented tissue component employing k-means segmentation algorithm. In the next step, they measured texture, shape, and color attributes of the images and segmented objects. In final step, they applied k-nearest neighborhood algorithm as a classification method to categorize images to two groups of cancerous and healthy.

Yu et al. [9] extracted several texture and shape features from histopathology images of non-small cell lung cancer and applied various classification methods to recognize cancerous tissues from the adjacent healthy ones and to distinguish lung adenocarcinoma from squamous cell carcinoma cancer types. Moreover, using a large set of relevant features, they designed a pipeline for predicting the prognosis rate of lung cancer patients to categorize between short-term survivors and long-term survivors.

1.4 Thesis objective

In this thesis, we primarily try to address the challenges associated with the improvement of automated techniques applied for analysis of histopathology images. We demonstrated the effectiveness of our designed pipeline for categorizing malignant and benignant cells presented in images obtained from lung adenocarcinoma and squamous cell carcinoma patients.

Artifact regions such as tissue folds and bubbles produce diagnostically unassociated measures within any feature extraction pipeline and thus reduce the efficiency of the classification system. Information about the location of tissue folds help to avoid selecting these areas for further processes.

Firstly, we searched for an optimized algorithm for recognizing the tissue folds regions within histopathology images and discard these areas from further processing while storing diagnostically relevant sections of images. Next, we used CellProfiler as a segmentation and feature extraction tool to derive various clinically important attributes from images. Finally, we utilized machine-learning algorithms to classify images to two categories of cancerous and healthy based on the extracted feature vectors.

The main novelty of this work is integrating the techniques developed for detection of the biologically relevant sections from high-resolution WSIs into the pipelines designed for characterizing malignant and healthy patients. Consequently, we designed a fully automated pipeline that achieves a significant improvement in accuracy compared to other studied models.

1.5 Thesis structure

The rest of the thesis is structured as follows. Chapter 2 reviews the theoretical background of the project; Section 2.1 provides an overview about machine learning methods and their applications for automating various engineering tasks. The literature relating to the biological background of the project is considered at section 2.2. The rest of chapter 2 discusses various parts of a supervised machine learning system applied in most computer vision problems. Chapter 3 presents the methodology of the work and describes various stages of the proposed automated pipeline and discusses the obtained results. Finally, the conclusions of the project and possible future works are discussed in chapter 4.

2. THEORETICAL BACKGROUND

2.1 Machine learning

Machine learning (ML) is a general concept referring to systems, which try to solve a set of prediction or detection tasks, based on learning from data. In fact, machine learning is a branch of computer science that utilizes pattern recognition theories and algorithms to produce a trainable computerized system. After learning, an ML system can perform the whole process of the desired pattern recognition tasks automatically without explicit programming [11].

ML algorithms enable analysis of the huge amount of quantitative data, which is otherwise impossible or time-consuming for human. Additionally, it is possible to assess the performance of an ML system quantitatively using different error metrics. As a result, ML systems have found numerous applications in different areas such as computer vision, communication systems, speech and text analysis, computational biology, health systems, and so on.

Based on the applied learning algorithm(s), ML systems are divided into three main categories of supervised, unsupervised, and reinforcement models. Most of the common ML models use supervised algorithms. It is also possible to combine two or all above methods to design a learning machine.

The task of a supervised system is to estimate the parameters of a defined pattern recognition model using labeled samples. Having a set labeled data $(\boldsymbol{x}_1, y_1)$, $(\boldsymbol{x}_2, y_2), ..., (\boldsymbol{x}_N, y_N)$ where $\boldsymbol{x}_k \in I\!R^P$ is the k-th example from the input space, and $y_k \in \{1, 2, ..., C\}$ is the corresponding label (target) from the output space, a supervised algorithm searches for the possible models that fit into the available data samples. Once the parameters of the model is known, it is possible to predict the output (label) of a query sample $\boldsymbol{x}_q$. The whole learning process in a supervised system includes three stages of training, validation, and testing. Accordingly, the labeled data is usually partitioned into three subsets. In the first step, the system feeds the set of training samples to an estimator to build a predictive model. In validation phase, the hyper-parameters of the model is tuned using validation set.

Finally, the performance of the fitted model is measured using the testing set. The efficiency of a supervised system for making correct predictions on the new unseen examples depends on how large is the number of labeled samples compared to the parameters of the predictive model.

However, in many learning situations, the available data set is not labeled. Unsupervised methods are systems that learns to extract patterns from a set of unlabeled data. Using the derived patterns, the model designs a function to describe some hidden behaviors of the system and make new predictions based on it.

Reinforcement systems learn through interacting with their environments. The learning procedure consists of a sequence of trials, which are replied by an error or a delayed reward to help the system to distinguish the desired situation.

Predictive models can be further categorized to two fields of classification and regression. Classification methods are used to predict discrete categories or target classes of an input object based on some defined or measured attributes of it. Whereas, in regression problems, the target is a continuous variable. The task of regression is to estimate a model, which is then used to find the output value of a new input variable.

2.2 Detecting cancerous patterns in histopathology images

Pathology studies the type and cause of a disease in the body tissue at cellular or molecular level using clinical or anatomical samples. A biopsy is a small sample of tissue or specimen, which is removed from the patient's body through surgery for further examination and analysis. In histopathology, biopsy samples are examined visually under the microscope to diagnose various tissue diseases.

Diseases like cancer change the usual appearance of tissues in cellular and molecular levels. Once the histopathology samples are provided, the pathologist observes the various slides of them under a microscope glass and searches for delicate patterns and signs of abnormalities. Therefore, the pathologist needs to investigate the fine details of available microscopic images carefully to find the indicators of malignant cells from the appearance of tissue slides.

The whole process includes three steps. Firstly, when the physician suspects that there are signs of malignancy or other tissue diseases in a patient, a small section of tissue is removed through surgery or biopsy for histopathology analysis.

In the next step, some techniques are applied on the tissue samples prior to placing

them under the microscope. The preprocessing includes three main stages of dehydration, freezing, and staining. The samples are dehydrated in consecutive steps to remove water from them. Next, tissue samples are either frozen or placed into a chemical fixative such as formalin to prevent tissue decay. Finally, the samples are sliced to thin layers and stained using one or more color pigments to help to detect different tissue components.

In the final step, a pathologist observes different sections of obtained slides carefully to examine various tissue patterns. Staining the histopathology slides aids to increases the contrast between different cell components. Therefore, staining techniques help the pathologist to distinguish various diagnosis patterns of the cells more efficiently.

One of the most common staining methods is to dye tissue slides with hematoxylin and eosin. In this technique, the nucleus of the cells are stained in blue or purple using hematoxylin, and cytoplasm and extracellular connective sections are stained in pink or red using eosin.

Cancer is the condition in the body when the cells grow and spread in an uncontrolled manner. Cancer can appear in any tissue of the body organs except hair, teeth, and fingernails. Cancer cells are recognized by some of their main characteristics such as abnormality, uncontrollability, and invasiveness [12].

Cancer cells are abnormal because they cannot function healthily like other cells and thus have no useful practice. Moreover, cancer cells divide in random manner, which leads to the formation of an unstructured mass. This is in contrast with the normal cells, which divide and grow in a regulated manner. Additionally, normal cells remain in their original organ, whereas malignant cells can move from one tissue to another. Cancer cells spread through body fluids such as blood or lymph circulations or by being directly implemented in the new organ.

The process of spread of malignant cells from an initial site to another organ is called metastasis and the secondary tumor is known as the metastatic tumor. The metastatic cells are similar to the cells of the primary site. Thus, cancer is always recognized by the primary organ it has started to grow [13]. For example, if a cancerous tumor has its primary site in the breast and later spreads to the liver, the tumor cells in the liver are cancerous breast cells and the tumor in the liver is called metastatic breast cancer (Figure 2.1). Detection of metastatic tumors is easier because the type of a cancer cell is mostly similar to the primary site where it is originated from and is different from the surrounding cells of the secondary site.

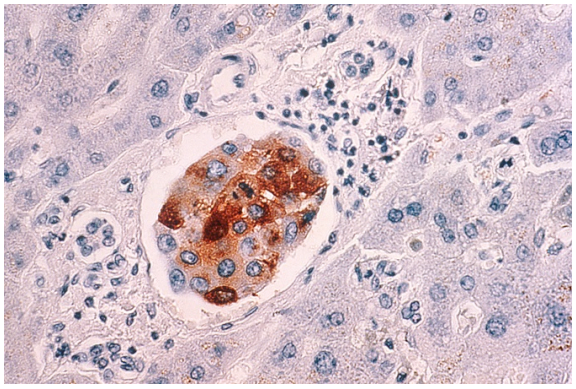


Figure 2.1 *Liver metastases from breast cancer [15].*

Diagnosis of cancer includes finding the accurate location of the origin of malignant cells as well as their types [12]. During the examination of a biopsy sample under a microscope, the pathologist investigates the different sections of a histopathology slide to find diagnostically relevant visual features. Some patterns are associated with detection of cancer if they are observed in specific tissue types. However, some other patterns are always considered as cancer indicators [13]. Figure 2.2 compares the tissue structure of normal and cancerous cells in histopathology images obtained from kidney, prostate, and the pancreas.

Malignancy indicators mainly include size, shape, color, density, and the arrangement of cells and nuclei of the cells. Figure 2.3 illustrates some of the differences between visual patterns of healthy and malignant cells.

The nuclei of the cancer cells are more active and contains more number of deoxyribonucleic acid (DNA) molecules. As a result, cancerous cells often have larger nuclei and their nuclei appears darker in stained tissue slides. In fact, high nucleus to cytoplasm ration (N:C) is an indicator of cancer [13].

In addition, the size and shape of the cells and nuclei of the cells are among the main attitudes for recognizing tumor sites. Cancer cells are often smaller or larger than surrounding normal cells. In fact, normal cells have specific sizes and shapes, which depend on the different functions of their corresponding tissues and organs. However, cancer cells neither have specific size nor function normally. In numerous cases, cancer cells appear as a cluster of cells with different sizes, without a clear boundary between them.[13]

Furthermore, normal cells follow a specific arrangement based on their types and functions. For example, inside breast, cells have glandular shapes. In contrast, breast cancer cells do not form glands or form glands of irregular shapes.

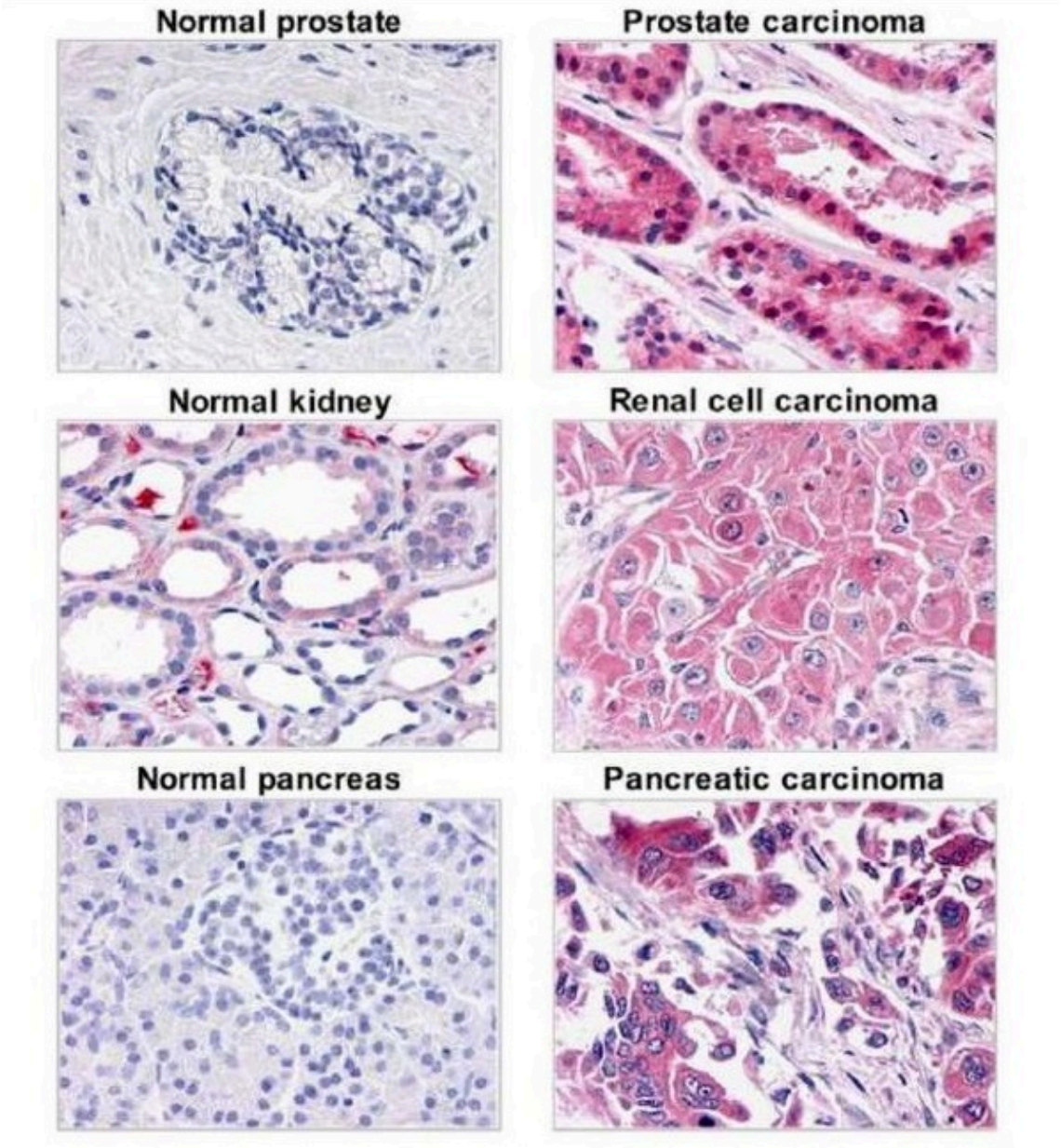


Figure 2.2 *Comparing appearance of cancerous and healthy tissues from three organs of kidney, prostate and pancreas [16].*

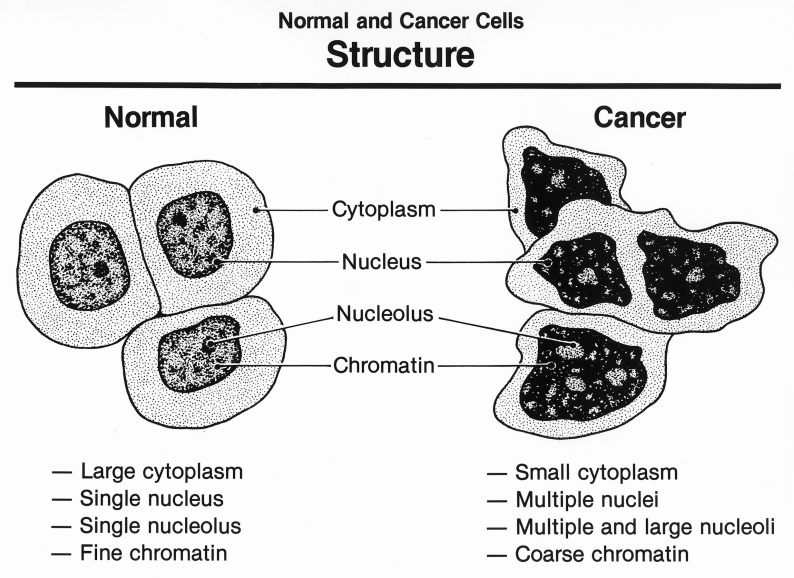


Figure 2.3 *Cancer vs normal cells. [14]*

2.3 Feature extraction

Features are a set of measured data or observed attributes, which are useful for discrimination purposes. Especially, when the objects themselves contain a large amount of redundant or non-informative information. Feature extraction is one of the main important parts of most supervised systems.

In the domain of image processing, feature extraction techniques aim to retrieve and understand visual content of images using different measurements. The majority of available strategies extract visual features using the image values in the spatial domain. However, some other methods use the data obtained through applying different transformation functions on images. For example, frequency components are employed in various content-based classification problems as they contain useful information about image edges.

In the spatial domain, image descriptors mainly include color and texture of the images and morphology of the image objects. Texture represents the statistical distribution of gray-level pixels in a defined block of the image. For example, the arrangement of pixel values in an image showing a field of crops is different from the one in the picture of a brick wall.

Texture features measure attributes such as homogeneity, contrast, and complexity of the images. Texture includes significant information about the structure and arrangement of objects with respect to each other and with respect to the surrounding environment [17].

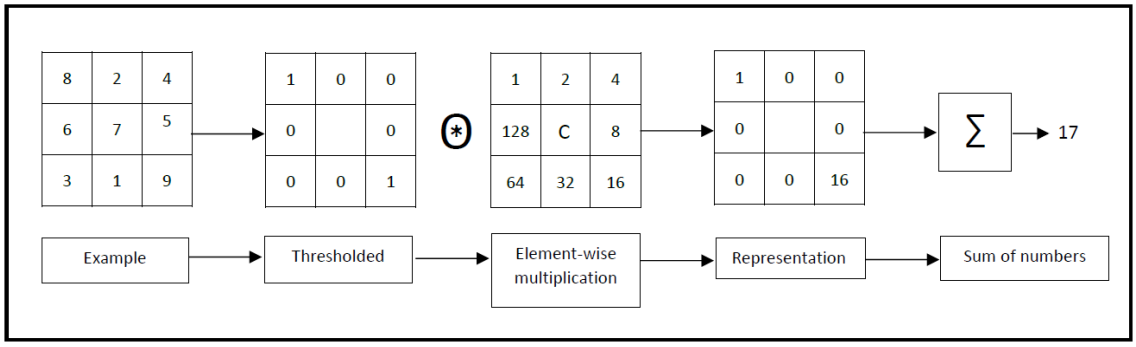


Figure 2.4 *An example of calculating LBP in an 8-neighborhood mask.*

Most of the texture extraction algorithms use mono-chromic images. For retrieving textural features from multi-channel pictures, a gray-scale conversion of images is used. It is also possible to extract information from each of the channels and concatenate them in a vector.

The statistical descriptors mainly include mean, median, variance, dispersion, average energy, entropy, skewness, and kurtosis. Statistical measurements that are obtained using single pixels are called "first order" features. Such descriptors can calculate simple attributes such as intensity histograms. Higher order operators like co-occurrence matrices, which use image blocks, can additionally measure pixel neighborhood relationships [19].

2.3.1 Local texture features

Local Binary Patterns (LBP) is one of the most powerful algorithms for describing textural characteristics of an image. It has been widely used in various computer vision applications from face recognition to cancer detection [20][7].

LBP algorithm labels each pixel of the image by comparing its value to the surrounding pixels in every local block of the image. Therefore, it transforms an image to an array of integer numbers. For measuring the label of each pixel, the original form of LBP considers a local window of 3 by 3. Thus, each pixel is compared with eight surrounding neighborhood pixels. However, in the generic form of LBP, the size of blocks varies based on the content of images and dimension of the target features.

The original form of LBP measures two different textural descriptors including a pattern and its strength. Figure 2.4 shows an example of the method used for calculating the label of the central pixel C in the original form of an LBP algorithm.

The model compares the value of each central pixel with its neighboring pixels and thresholds them based on this comparison as

$$x = \begin{cases} 1, & x \geq C \\ 0, & otherwise \end{cases}, \tag{2.1}$$

where x is the intensity value of the surrounding pixel and C is the intensity value of the central pixel. Next, the thresholded values are multiplied by a weight matrix with 8-bit numbers and summed to build a label for the central pixel. The contrast value is obtained by subtracting the average of pixels greater than the central pixel from the average of pixels smaller than the central pixel. The combination of two values (i.e. LBP and Contrast) is used as a two dimensional vector for describing texture of the image [21].

Local phase quantization (LPQ) is a blur-invariant descriptor, which aims to detect contrast information of the images. Fourier transformation function, transfers images from spatial domain to frequency domain. In frequency domain, image components are phase and magnitude. Magnitude values are the main data used in image enhancement tasks such as noise removal algorithms. However, phase components carry information about image edges and contours. Thus, compared to frequency magnitudes, local phases contain more data about the image texture.

Proposed in 2008 by Ojansivu and Heikkilä [22], LPQ is a robust to the method of image blurring. It computes local phase information of low-frequency components over a rectangular neighborhood at every pixel position. In the next step, the phases are quantized in an eight-dimensional space. Finally, the resulting code words are presented in the form of a histogram to describe texture of the images.

Gray-level co-occurrence matrix (GLCM) measures the distribution of co-occurring pixel values at a given offset. It is obtained from a gray scale image and represents the angular and distance relationships of the adjusting pixels in the spatial domain [17] .

In fact, GLCM measures how often a pixel relationship (such as 0-0, 0-1, 1-0, and 1-1 in a binary image) occurs in any specific angular direction. The four main directions of adjacency are horizontal, vertical, top left to bottom right, and top right to bottom left (Figure 2.5).

GLCM is a square matrix and its dimension is equal to the number of gray levels presented in the image. Matrix elements are obtained by counting the number of

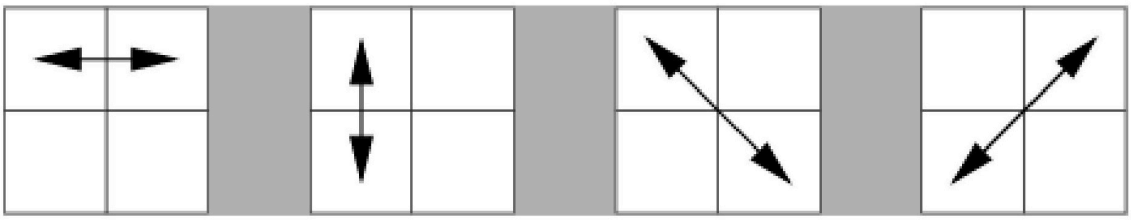


Figure 2.5 *The four directions of adjacency for calculating the co-occurrence matrix*

times any adjacent pixel values are occurring with respect to a given direction. Finally, the matrix is divided by the total number of registered counts to generate the probability of different occurrence rates. For any defined direction, the GLCM can be written as:

$$GLCM = \begin{bmatrix} p(1,1) & p(1,2) & ... & p(1,N_g) \\ p(2,1) & p(2,2) & ... & p(2,N_g) \\ \vdots & & & \\ p(N_g,1) & p(N_g,2) & ... & p(N_g,N_g) \end{bmatrix} \tag{2.2}$$

In which $p(i,j)$ is a count for the number of times that "i" and "j" are occurring with respect to a specified direction. For example, if the angular direction is considered as the top right to the bottom left, then the matrix elements are the number of times in which $f(x,y) = i$ and $f(x+1, y+1) = j$. The parameter of N_g indicates range of pixel values.

Once co-occurrence matrix is calculated, it is possible to extract various statistical quantities of images from them. For instance, Haralick features are a set of measurements extracted from GLCM. They represent several textural attributes such as contrast, correlation, and entropy [18].

2.3.2 Gabor filters

Gabor filters are one of the most effective methods for detection of textural features. They are used in a wide range of applications such as text processing and face recognition problems. In fact, one of the main advantages of Gabor filters is that their functionality is similar to the perception mechanism of the human visual system [24].

Gabor filters are a set of filters applied in different directions and magnitudes to the images in frequency domain [25]. Thus, prior to extract Gabor features, images

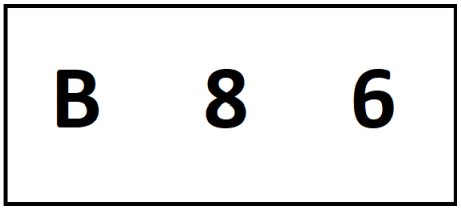


Figure 2.6 *An example of shapes separable by their Euler numbers*

are converted to frequency domain using a Fourier transformation function. In the next step, variations of pixel values in different orientations and scales are detected using a bank of Gabor filters. These filters can retrieve image correlation bands in different orientations. For example, objects with horizontal textural lines such as window blinds have higher scores in horizontal direction.

2.3.3 Shape descriptors

Shape descriptors are a category of attributes extracted from gray-scale images to represent different visual features of the image objects such as solidity, convexity, Euler number, eccentricity, circularity ration, profiles, and least inertia. The two main algorithms used in shape feature extraction techniques are counter-based and region-based models. Counter-base methods searches for the pixels located on the boundary of objects, whereas region-based algorithms exploit shape parameters considering all pixels within the shape region. [27][26]

- Circularity ratio

  Circularity ratio defines how similar is a shape to a circle using measurements from the area, the perimeter, and mean and variances of the distribution of radial distances.

- Center of Gravity

  The gravity center (or centroid) of a shape is the location of its center of mass. Centroid is invariant to the various distributions of the boundary points. It helps to find the location of objects in 2-D and 3-D images and their distances with respect to each other.

- Eccentricity

  Eccentricity is the ratio between major axis and the minor axis of a shape. It is a useful attribute for distinguishing between different ellipse-like shapes such as nuclei of a cell. After defining the boundary of cell components, various

mathematical models are applied to predict the cell shape parameters such as length and direction of their major and minor axises [28]. Eccentricity of an ellipse ranges from zero to 1, with zero demonstrating a circle and 1 showing a parabola.

- Solidity

  Solidity defines to which extent an object is convex or concave. solidity is defined as:

$$Solidity = \frac{A_s}{H}, \tag{2.3}$$

  whereas A_s is the whole pixel area of the object and H is the area of the convex envelope of the shape. Convex envelope (also called convex hull) of a shape is the smallest convex set of points in euclidean plane which covers that shape.

- Euler number

  Euler number is an effective topological descriptor ratio describing continuity of the boundary of shapes. It is defined as the number of connected component minus the number of holes in border regions of the object. Euler number is a useful feature to separate simple shapes with different boundary distributions. Figure 2.6 shows some examples of such patterns.

2.4 Classification methods

2.4.1 Linear discriminant analysis

Linear discriminant analysis (LDA) is a member of the family of linear predictive models. Linear methods are applicable for problems in which the data are separable using linear decision boundaries. Such a boundary divides feature space into different class regions.

proposed originally by R.A. Fisher in 1936 [29], the LDA algorithm classifies objects by reducing the dimensionality of the given data vectors through finding a linear combination of them. Having a set of labeled data in two or more-dimensional feature space, the LDA method firstly reduces the size of the feature space by projecting data onto a line. Next, LDA algorithm classifies objects based on their location on the projection line. The problem is to find the direction of the projection line such that the data are best separable after mapping to this line.[30][31]

For a binary classification problem, LDA tries to find a linear combination of different features that can separate the two classes in the best possible way. Given the training data vector $\boldsymbol{x} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots]$ and label vector $\boldsymbol{y} = [y_1, y_2, \cdots]$, the problem of LDA algorithm is to estimate the direction of projection line $\boldsymbol{w}$ by maximizing separability score $J(\boldsymbol{w})$, which is defined as

$$J(\boldsymbol{w}) = \frac{\left(\boldsymbol{w}^T \boldsymbol{\mu}_1 - \boldsymbol{w}^T \boldsymbol{\mu}_2\right)^2}{\boldsymbol{w}^T \boldsymbol{\Sigma}_1 \boldsymbol{w} + \boldsymbol{w}^T \boldsymbol{\Sigma}_2 \boldsymbol{w}}, \tag{2.4}$$

where μ_1 and μ_2 are the mean values of the distributions of class 1 and class 2 on the projected line $\boldsymbol{w}$, respectively, and, $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are the corresponding covariance matrices of the two classes. $J(\boldsymbol{w})$ can be written as

$$J(\boldsymbol{w}) = \frac{\boldsymbol{w}^T \boldsymbol{S}_B \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{S}_W \boldsymbol{w}}, \tag{2.5}$$

where $\boldsymbol{S}_B$ is the distance between class means on the projection line and $\boldsymbol{S}_W$ is the sum of class variances:

$$\boldsymbol{S}_B = \left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)^T, \tag{2.6}$$

$$\boldsymbol{S}_W = \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2. \tag{2.7}$$

The problem of finding the direction of projection line W is equal to finding eigenvalue λ and the corresponding eigenvector $\boldsymbol{w}$ from

$$\boldsymbol{S}_B \boldsymbol{w} = \lambda \boldsymbol{S}_W \boldsymbol{w}. \tag{2.8}$$

The two classes are best separable on the projection line if the distance between the mean values of their distribution functions will be maximized. However, to achieve the best classification result, it is also important to minimize within-class variances. Therefore, the best projection line provides the least overlapping between two classes by maximizing the distances of the mean values and minimizing within-class variances of each of the populations, simultaneously.

Figure 2.7 shows the distributions of a set of binary data vectors in red and blue

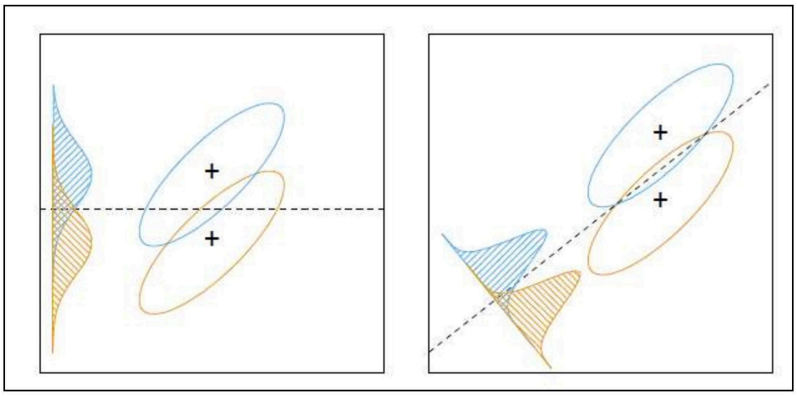


Figure 2.7 *LDA: Projection of a binary sample on two different directions[31]*

Day	Temperature	Outlook	Wind	Humidity	Target Decision
1	High	Rain	High	High	No
2	High	Sunny	Low	Moderate	No
3	Moderate	Sunny	Low	Moderate	Yes
4	Moderate	Sunny	High	Moderate	No
5	Moderate	Sunny	Low	High	Yes
6	Moderate	Rain	Low	High	No
7	High	Sunny	Low	High	Yes

Figure 2.8 *Observations for weather features and corresponding decisions on riding a bicycle during seven days.*

colors, respectively. The left and right pictures show the mapping of the data on two different projection lines. The distributions of the data is different on each of the projection lines. As it can be seen from the figure, the distance between the mean values is larger in the left picture. However, the projection line in right produces a better classification by reducing the amount of overlapping data through minimizing the class variances.

2.4.2 Decision trees

Decision tree (DT) is a non-parametric supervised learning classification and regression algorithm, which is based on calculating conditional probabilities. Considering a set of conditions as decision rules, one can find a sequence of attribute values, which leads to a specific decision target. As an example, the target action can be a binary decision on riding a bicycle based on different weather attributes.

A Classification tree is a tree-like graph in which leaves (end nodes) indicates class labels (target variables) and non-leaf nodes (interior nodes) represent a decision rule

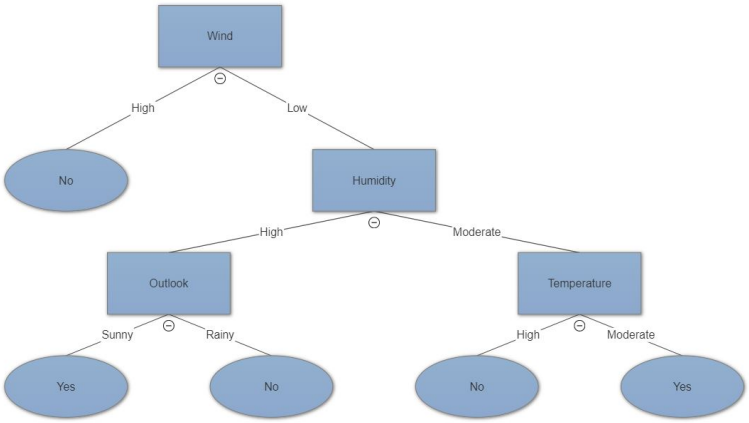


Figure 2.9 *An example decision tree obtained for riding a bicycle*

(condition). Drawing a decision tree also helps to represent the decision-making procedure visually.[32]

As an example, Figure 2.8 shows a set of different observations (training samples) that influence a decision on riding a bicycle. Figure 2.9 shows the corresponding decision tree obtained for such observations. For several presented attributes such as weather temperature or outlook, it is important to decide which feature has more critical impact on the final classification. A decision tree is constructed by arranging the different attributes from root to leaves based on their importance.

Pure attributes are the features that solely separate the two classes. In above example, "High wind" is a completely pure node which leads to a decision of "not riding a bicycle" in all observations.

One of the main problems of classification trees is over-fitting, which occurs when the number of recorded observations is too less compared to the number of involved attributes. This situation leads to forming of complex trees, which tend to memorize training samples and have less capability to generalize to the new unseen data. Strategies such as pruning of the tree helps to avoid over-fitting [32]. Pruning includes removing some branches than contain less important attributes.

Additionally, Bagging technique is one of the methods used for training decision trees. This method selects a random sample of training data in a recursive manner, and in each iteration, learns a tree with the selected samples. The result of a new prediction is calculated by averaging output values of several trained trees in regression tasks or by taking majority vote in classification problems.

2.4.3 Random forest

Random forest (RF) is a type of decision tree that tries to avoid over-fitting problem happens in normal decision trees by randomly splitting the data and features to smaller subsets. As its name indicates, instead of using all input data and all given features in one single tree, RF algorithm tries to build a set of smaller imperfect trees.

As mentioned in the previous section, using all training data and attributes in one single tree might force the algorithm to fit to irregular and noisy patterns. This leads to the situation of over-fitting where the model is memorizing the training set and has a weak ability for generalizing to unseen test data.

The goal of RF algorithm is to reduce the variance of a DT by averaging outcomes of several decision trees. [33]. Therefore, numerous of imperfect trees are trained by using a smaller portion of training data and feature vectors. The number of trees can change from tens to thousands.

Once many imperfect trees are trained, the final prediction is obtained using "majority-vote" strategy and the decision labels are calculated based on the portion of trees that vote for each class.

In an RF classifier, various combinations of features can be selected using different strategies. For example, in the shuffling method, the performance of the trees trained using different sets of attributes are tested using accuracy score. The algorithm obtains a score for the effectiveness of any set of features by comparing accuracies of the trees obtained using randomly shuffled attributes. Omitting an important feature will decrease the accuracy dramatically. In contrast, accuracy remains almost unchanged by shuffling non-important features.[34]

The selection of the feature is usually done in a without replacement sampling manner meaning that it is not allowed to have any of the same feature sets in two different trees. However, the selection of an appropriate set of training data is usually done using with replacement manner. With replacement refers to sampling strategy where a member might be presented in different samples. This method aims to avoid producing correlated trees. [33]

2.5 Error metrics

Error metrics are defined to measure the performance of a machine learning system and to evaluate the capability of the fitted model to generalize to the new unseen

		Predicted classes	
		Positive	Negative
Actual classes	Positive	True Positive	False Negative
	Negative	False Positive	True Negative

Figure 2.10 *Confusion matrix in a binary classification problem [35]*

data. In a supervised learning method, the whole labeled data set is usually divided to three sets of training, validation, and testing. The parameters of the classification model is initially fitted using the training set. Next, the hyper-parameters of the model are adjusted using validation dataset. Finally, the performance of the trained model is evaluated using test dataset.

Various performance assessment techniques and formulas have been developed in literature. However, it is not possible to determine which assessment method is the most efficient way to score the result of all classification systems. In general, selecting one or a set of specific error metrics depends on the nature of the problem and the characteristics of the data. Thus, for any specific collection of data and defined problem(s), it is important to select one or a set of appropriate evaluation metrics. [35]

2.5.1 Confusion matrix

Confusion matrix or error matrix is a method to visualize the performance of a classification system. Confusion matrix is mostly used for supervised systems in which the correct labels of testing samples are available. The output of the classification system is the predicted labels of the testing samples. Therefore, it is often convenient to draw a matrix to compare how many of the predicted labels of each class is a true prediction and how many is a false prediction [35]. A schematic of a confusion matrix for a binary classifier is shown in figure 2.10 .

An example of a binary classification task is to divide a group of people to two classes of "healthy" and "sick" based on the results of a medical test. After applying the test, the results can be categorized in two groups of positive and negative, indicating "sick" and "healthy" people, respectively. True and False refers to the correct predictions and incorrect predictions, respectively.

Therefore, True Positive (TP) shows the number of samples from positive class, which are correctly predicted as positive (i.e. the number of sick people who are

Table 2.1 *Error metrics derived from confusion matrix [35]*

Error metric	Definition
Accuracy	TP+TN / (All predictions)
Sensitivity: TP rate	TP/P=TP/(TP+FN)
Specificity: TN rate	TN/N=TN/(TN+FP)
Fall out: FP rate	FP/N= FP/(FP+TN)
Precision: PPV	TP/(TP+FP)
Negative Predictive Value: NPV	TN/(TN+FN)

detected as sick). Similarly, True Negative (TN) indicates the number of negative samples that are correctly labeled as negative by the classification system (i.e. the number of healthy people who are detected as healthy).

In contrast, false alarms correspond to incorrect detections. In the example of medical test, FN is equal to the number of sick people who are detected as healthy and FP is the number of healthy people who are detected as sick.

For an ideal classifier, the number of false alarms is equal to zero, which shows that the system is detecting true labels for all the samples in both positive and negative classes.

2.5.2 Error metrics derived from confusion matrix

Various error metrics can be derived from confusion matrix. Each of these metrics can examine one or more aspects in the performance of a classifier. Table 2.1 illustrates some of the most important error metrics obtained using elements of a confusion matrix [35].

"Accuracy" is the ratio of correct predictions to the total predictions made by the classifier. "Accuracy" is one the main applied assessment measures in some classification problems. However, it is not very useful metric in many of the practical situations as it combines true positive and true negative rates. As a result, "accuracy" is not a suitable metric for a set of unbalanced data.

In contrast, "sensitivity" and "specificity" are often more informative since they measure the rate of correct classifications for true and false categories, separately.

"Sensitivity" or True Positive rate is the portion of positive samples that are truly detected as positives. For example, the number of people who are detected as sick to the total number of people who are truly sick. Total sick numbers are the sum of TP (sick and recognized as sick) and FN (sick and not recognized as sick). High

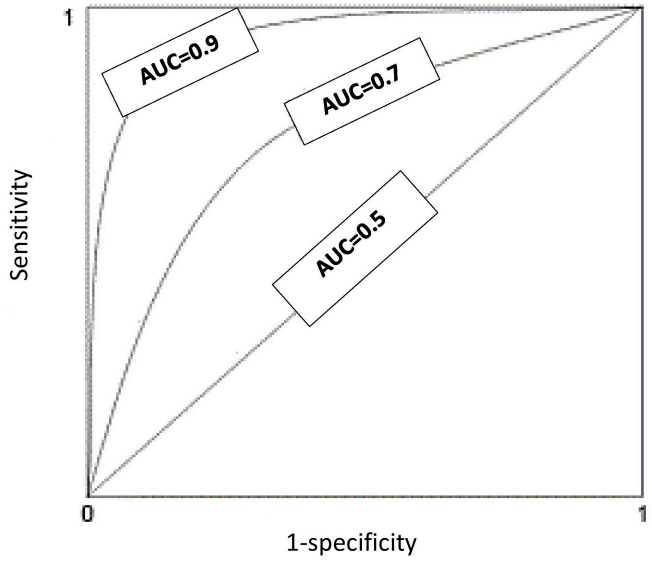


Figure 2.11 *ROC curve*

"sensitivity" shows that the classification system can detect sick people good.

Accordingly, "specificity" or True Negative rate is the portion of negative samples that are correctly classified as negatives. For example, the portion of healthy people that are detected as healthy to the total number of healthy people. High "specificity" indicates that the system is not producing too many false alarms.

FP rate is the ratio of all negative samples (TN + FP) that are incorrectly classified as positives (FP). For example, the ratio of healthy people that are wrongly classified as sick. FP rate is equal to 1-Specificity.

"Precision" or "positive predictive value" is the ratio of the positive samples, which are detected as positives (TP) to the whole number of samples that are predicted as positive (TP + FP).

It is also notable that both precision and sensitivity are measuring different quantities related to the prediction of positive class members. While sensitivity indicates that how much a classification system is sensitive about detecting positive samples, precision shows how much it is precise about this task.

2.5.3 Receiver operator characteristic

Receiver operator characteristic (ROC) curve (Figure 2.11) is another error measurement metric, which helps to visualize the performance of a classification system in form of a curve. Particularly, it demonstrates how predicted values are relevant compared to true values.

The ROC curve is an important tool to evaluate the performance of different binary classifiers in many Data Mining problems. The curve is obtained by drawing TP rate (sensitivity) versus FPR (1-specificity) for different threshold values, which is used to separate positive classes from negative ones.[35]

The performance of the classifier is better if the curve is close to the top-left corner. A random classification has a diagonal ROC curve drawing between (0,0) to (1,1). In multi-class problems, it is possible to build a ROC curve to measure the performance of the classifier in a one-versus-all or a one-versus-one approach.

The ROC curve is not very representative for many real-world problems in which the whole data set is unbalanced. However, even for unbalanced data, the area under the curve (AUC) can be used as a suitable measuring tool to assess and compare the performance of different classifiers.

2.5.4 Cross validation

Cross validation (CV) is a method to estimate the performance of the predictive systems on unseen data. Indeed, it is a strategy to validate how well a classifier, which is trained using labeled data, can generalize to new unseen data.

As discussed, one method of testing the performance of a predictive model is to split the data randomly to the training and testing sets. In usual cases, a higher portion of data is assigned to the training set and a lower portion is kept for the testing phase. For example, 80 percent is considered for training and 20 percent is used for testing.

However, random splitting of data to two fixed sets of training and testing has some disadvantages. Firstly, it is not an efficient method for unbalanced data. Secondly, it is not a suitable technique for small datasets, where an enough number of labeled data is not available. Training with a small number of examples might lead to the problem of over-fitting in which the model memorizes the training set. Such a model fits well to training samples, but cannot generalize to the testing data. Therefore,

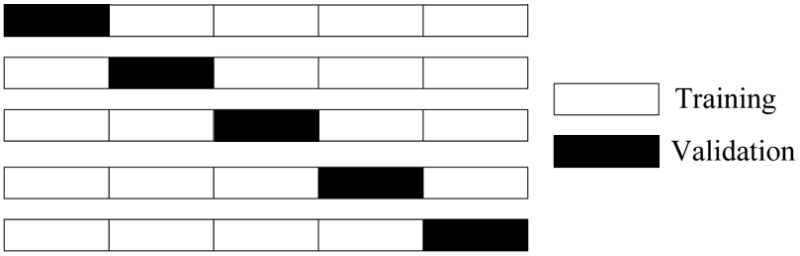


Figure 2.12 *Cross validation: One iteration of a five-fold cross validation*

a fixed split of data to training and testing set does not ensure achieving the best possible solution and parameters for the trained model.

In the cross validation method, the whole dataset is partitioned to complementary sub-sets in an iterative manner. In each round, a portion of data is used as labeled data for fitting the parameters of the predictive model and the remaining portion is used for evaluating the performance of the fitted model. In fact, a validation set is one split of the labeled data, which is not used for training. Instead, it is used for testing how well the parameters of the fitted model are generalizing to new data. [36]

In a k-fold cross validation strategy, the data is divided to k partitions. Therefore, the classifier is trained k times iteratively. In each repetition, the classifier is trained using k-1 splits of the data and the performance of the trained classifier is tested using the left partition. An example of one whole iteration in a five-fold cross-validation strategy is shown in Figure 2.12.

2.6 Dealing with unbalanced data

Mining unbalanced data in both training and evaluation task is one of the main challenges in data engineering. The data set is considered unbalanced if instead of having a set of almost equal samples for all available categories, some classes include majority samples and other classes have minority samples. [37]

As an example, in the task of anomaly detection in wireless communication systems, the set of training data are mostly collected from normal conditions (major class) and a very small set of samples are available from fault signals (minor class).

One of the solutions of this problem is to balance data artificially. For example, by assigning a larger weight to the smaller class. Alternatively, it is possible to over-sample the lower populated class or under-sample over-populated class. Over-sampling strategy balances the size of data by increasing the size of samples from

the less frequent class using some techniques like repetition or bootstrapping. In contrast, under-sampling balances the data by decreasing the size of samples from abundant class. [34].

Additionally, it is possible to use some specific validation strategies while dealing with unbalanced data. For example, the technique of "stratified" cross-validation leaves the classes unbalanced but keeps the original proportion of the classes while assigning them to each fold. [37]

3. EXPERIMENTAL SETUP AND RESULTS

The goal of this section is to provide a model to distinguish between benign (healthy) and malignant (cancerous) tissues in histopathology images of lung adenocarcinoma and squamous cell carcinoma samples. The suggested methodology consists of three main stages of preprocessing, feature extraction, and classification. In the first step, some pre-processing tasks were applied to the whole-slide images to select clean and informative patches from each picture. The selected patches were cropped and stored as a new data set for further processing. In the next step, cells and nuclei objects were segmented and biologically informative features were measured from images and segmented objects. The results of feature extraction stage were stored in the form of data vectors. Finally, different machine-learning algorithms were used to classify whole-slide images to two groups of healthy and cancerous based on the extracted features. A schematic of the proposed pipeline is illustrated in Figure 1.

3.1 Image dataset and data collection

In this step, 1067 histopathology images of lung adenocarcinoma [38] and 1060 images of squamous cell carcinoma [39] were downloaded from the legacy archive of The Cancer Genome Atlas (TCGA) [40]. All files belong to the open-source category and were downloaded from clinical data available at Genomic Data Commons (GDC) portal of National Cancer Institute (NCI) [41]. GDC provides a united data repository for studies in cancer genome.

All downloaded images were high-resolution pictures with size of tens of gigabytes. The image files had ".svs" extension and contained meta-data. The files were accessed using Bio-Formats library [42] in Matlab [43]. "Bio-Formats" is a software developed mainly for reading and writing pixel values and metadata of the common file formats used for life science images [44].

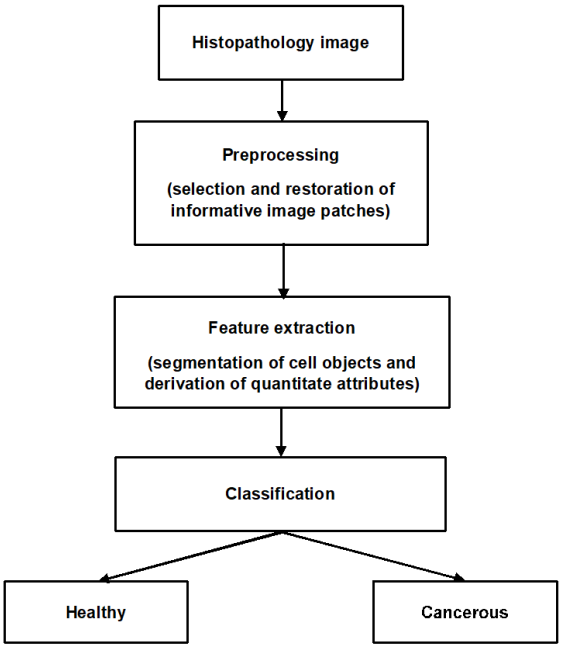


Figure 3.1 *Pipeline of the applied methodology for automated detection of cancer from histopathology images*

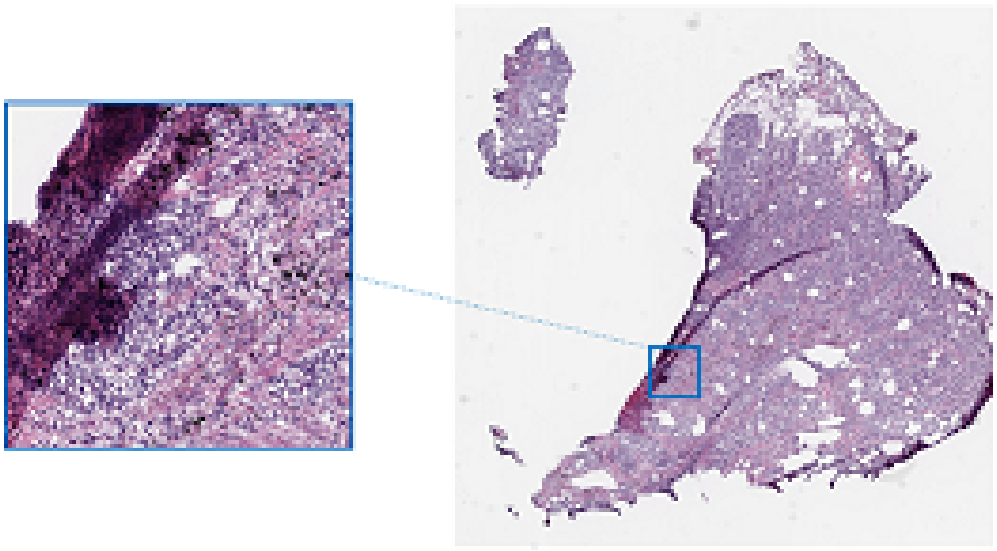


Figure 3.2 *An example of a whole-slide histopathology image obtained from lung adenocarinoma patient (right) and a cropped 1000 * 1000 patch affected by tissue folds.*

3.2 Preprocessing of data

3.2.1 Problem statement

Histopathology images used in this work were high-resolution pictures with typical dimensions ranging from 5000×5000 pixels to 50000×50000 pixels and higher. However, using the whole slide images containing tens of thousands of pixel values would produce a redundant and laborious task for any classification system. Furthermore, in most whole-slide histopathology images a large section of the picture is covered by the areas that are not useful for the task of cancer detection such as tissue background or sparse regions. Therefore, it is important to develop an algorithm which would be able to distinguish informative areas of the whole-slide images. Such algorithm can be applied in pre-processing step to select diagnostically relevant regions of the images and discard non-relevant areas from further processing.

However, detecting instructive regions is a non-trivial task because of complicated texture and high structural diversity of histopathology images. For detecting and selecting informative patches of whole-slide images, one need to search for the image attributes that indicate presence of cancerous cells in the tissue. Such attributes can demonstrate diagnostically important regions of the images.

As discussed in 2.2, the cancerous cells grow in an uncontrolled manner and mostly form dense clusters of arbitrary shapes. Therefore, the densest areas of images are usually more informative than the other sections as they are more likely to include malignant cells. However, some image artifacts such as areas containing tissue folds have also high pixel densities. Nevertheless, if the selection of relevant patches would occur solely based on the density of the patch, it will result in detection of many tissue folds regions. Thus, algorithms that are more intelligent are needed for the pre-processing step to detect instructive areas of the whole slide images while avoiding tissue folds simultaneously.

Figure 3.2 illustrates an example of whole-slide histopathology images (right) and a randomly selected patch of size 1000×1000 pixels (left) that is mostly covered by tissue folds. Furthermore, the dataset contains both cancerous images as well as images obtained from adjacent healthy tissues. Thus, we needed to search for a unique and effective pre-processing method which could be applied on both healthy and cancerous groups.

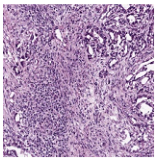
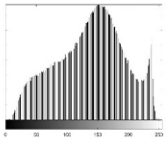
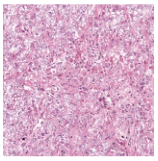
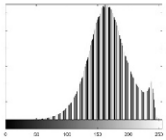
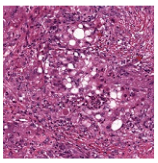
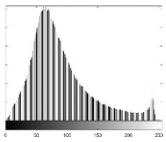
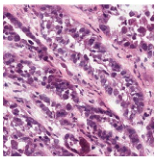
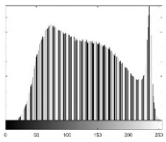
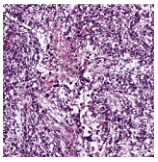
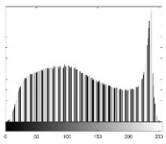
Image patch (1000 *1000)	Histogram of the green channel	Mean	Skewness	Kurtosis	Variance of tiles averages	Variance of tiles maximums	s/v
		137	-0.18	2.30	61.46	1.72e+03	0.00
		167	-0.01	3.05	14.62	35.31	0.00
		89	1.16	4.13	51.07	1.90e+03	0.00
		132	0.23	1.98	95.08	6.02e+03	0.00
		128	0.16	1.82	77.98	4.13e+03	0.00

Figure 3.3 *Example of some cropped 1000 * 1000 patches labeled as diagnostically relevant regions and their statistical measurements*

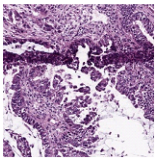
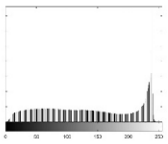
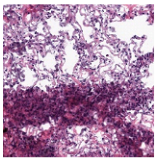
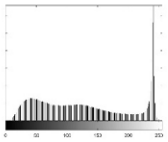
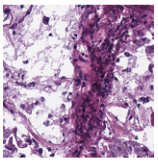
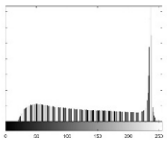
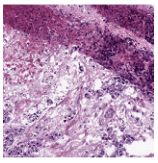
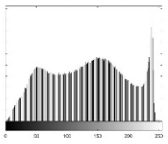
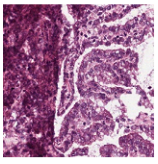
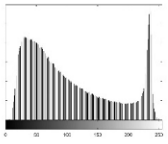
Image patch (1000 *1000)	Histogram of the green channel	Mean	Skewness	Kurtosis	Variance of tiles averages	Variance of tiles maximums	s/v
		137	-0.03	1.63	1.15e+03	4.77e+03	0.04
		126	0.27	1.73	1.99e+03	8.08e+03	0.11
		139	0.03	1.58	1.92e+03	8.34e+03	0.08
		132	-0.01	1.95	1.59e+03	7.05e+03	0.08
		104	0.65	2.12	1.33e+03	8.99e+03	0.12

Figure 3.4 *Example of some cropped 1000 * 1000 patches containing tissue folds and their statistical measurements*

3.2.2 Methods

In pre-processing step we focused on designing an algorithm to detect 10 relevant patches of size 1000×1000 from each histopathology image. For this purpose, we developed and designed a machine learning system, which can decide if any randomly selected patch in a whole-slide picture is chosen from an informative section. After designing and training such a classifier, it can be applied to the whole data set to detect 10 informative patches of each image and store them for further processing.

In order to design a classification system for preprocessing step, we needed a labeled database containing informative and non-informative sections of histopathology images of our data set. To produce such a labeled dataset, we randomly selected 20 whole-slide histopathology images from each cancer type. These images were discarded from the data and assigned to the preprocessing step. After providing a labeled data set, below steps were followed for each cancer type separately.

Firstly, background regions of the selected images were segmented using simple thresholding methods, which work based on the pixel density of the patches. Next, 100 patches of size 1000×1000 were selected randomly from each of the 20 images presented in the assigned dataset. Thus, for each cancer type, we gathered $100 = 2000$ patches from non-background areas of histopathology images.

Finally, the selected patches were annotated manually as relevant or irrelevant based on their appearance. Patches showing areas of the images with high intensity and homogeneous texture were selected as relevant regions. In contrast, image patches containing areas of tissue folds or sparse regions were labeled as irrelevant.

In the next step, we needed to search for the most relevant attributes of the image patches that can be used to distinguish between relevant and non-relevant regions. Statistical characteristics of an image have been used in literature for detecting diagnostically relevant areas of the images [3][4]. In this work, we measured various statistical attributes to detect and discard tissue fold areas as well as sparse regions. The main applied features include mean, median, variance, skewness and curtsies of the histogram of the patches.

As histopathology images were stained in blue and red using HE method, the green channel can present behavior of the intensity values purely, excluding staining characteristic. Thus, statistical features were measured using information of the green channel.

Moreover, tissue folds are highly saturated in color. Thus, the patches containing

tissue folds can be detected by their high saturation to intensity values [6][5]. Therefore, the mean value of saturation to intensity measurements from each patch was calculated and added as an element to the feature set to increase efficiency of the system for detecting tissue folds regions.

Additionally, we splitted each 1000×1000 patch to the smaller non-overlapping tiles of size 200×200 and calculated the maximum intensity values of each tile. Next, we measured the variance of maximum values of the tiles inside a patch. This strategy helps to detect intensity variations inside a patch in a larger neighborhood. Thus, it enables the system to detect inhomogeneous patches which were partly covered by tissue-folds or sparse regions. These types of patches were considered as non-informative regions in our labeled dataset.

Some examples of patches labeled as relevant as well as patches selected from tissue fold regions (irrelevant) are shown in Figure 3.3 and 3.4, respectively. The Figures also compare the main statistics derived from the green channel of the images in two above-mentioned categories. The last columns show the saturation to intensity values measured using information of saturation and intensity channels. As it can be seen from the pictures, the patches that are partly covered by tissue folds have greater variations in the average and maximum tile values. Moreover, tissue fold regions are more saturated in color and have higher saturation to intensity values compared to non-fold areas. The similar study confirms these results for the most image patches presented in the labeled dataset.

To determine if the statistical features can truly distinguish tissue folds and sparse regions of the images from diagnostically relevant areas, we fed them to a classification system and measured its performance using different error metrics. For this purpose, the statistical features were measured for all the data presented in our labeled dataset. Next, the data were divided to two groups of training and testing. The features vectors obtained from statistical measurements were used to train a classifier for the task of distinguishing between relevant and irrelevant image patches. Finally, different assessment tools were applied to test the performance of the system on the test data.

3.2.3 Results

To determine if statistical features can distinguish between relevant and non-relevant images patches, we used three different classifiers including LDA, SVM with Gaussian kernels, and RF with 200 trees. To measure the performance of the applied classifiers, we separated the whole data to two sets of training and testing using

Table 3.1 *Comparison of various classification models for detecting diagnostically relevant patches from lung adenocarcinoma images*

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average	Variance
Accuracy							
LDA	0.82	0.85	0.82	0.85	0.83	0.83	2.7e-4
SVM	0.86	0.89	0.88	0.87	0.86	0.87	1.2e-4
RF	0.89	0.87	0.89	0.89	0.87	**0.88**	9.2e-5
Sensitivity							
LDA	0.91	0.91	0.92	0.91	0.93	**0.91**	6.6e-5
SVM	0.88	0.86	0.89	0.85	0.87	0.87	3.2e-4
RF	0.91	0.88	0.92	0.88	0.87	0.89	4.2e-4
Specificity							
LDA	0.73	0.8	0.72	0.78	0.73	0.75	1.3e-3
SVM	0.85	0.91	0.87	0.89	0.85	**0.87**	8.2e-4
RF	0.87	0.86	0.86	0.9	0.87	**0.87**	2.7e-4
Precision							
LDA	0.76	0.83	0.77	0.81	0.79	0.79	8.0e-4
SVM	0.84	0.91	0.88	0.89	0.86	**0.88**	7.6e-4
RF	0.86	0.87	0.87	0.9	0.88	**0.88**	1.8e-4

Table 3.2 *Comparison of various classification models for detecting diagnostically relevant patches from lung squamous cell carcinoma images*

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average	Variance
Accuracy							
LDA	0.8	0.82	0.71	0.8	0.77	0.78	1.7e-3
SVM	0.82	0.85	0.8	0.88	0.81	0.83	8.5e-4
RF	0.86	0.85	0.82	0.87	0.84	**0.85**	3.9e-4
Sensitivity							
LDA	0.95	0.94	0.89	0.93	0.95	**0.93**	6.1e-4
SVM	0.89	0.87	0.83	0.93	0.89	0.88	1.3e-3
RF	0.92	0.84	0.84	0.92	0.89	0.88	1.6e-3
Specificity							
LDA	0.62	0.67	0.53	0.66	0.56	0.61	3.8e-3
SVM	0.74	0.82	0.77	0.81	0.73	0.77	1.7e-3
RF	0.79	0.85	0.8	0.81	0.78	**0.81**	8.0e-4
Precision							
LDA	0.74	0.78	0.66	0.74	0.7	0.73	2.0e-3
SVM	0.8	0.86	0.79	0.84	0.78	0.81	1.2e-3
RF	0.83	0.88	0.81	0.84	0.82	**0.84**	7.3e-4

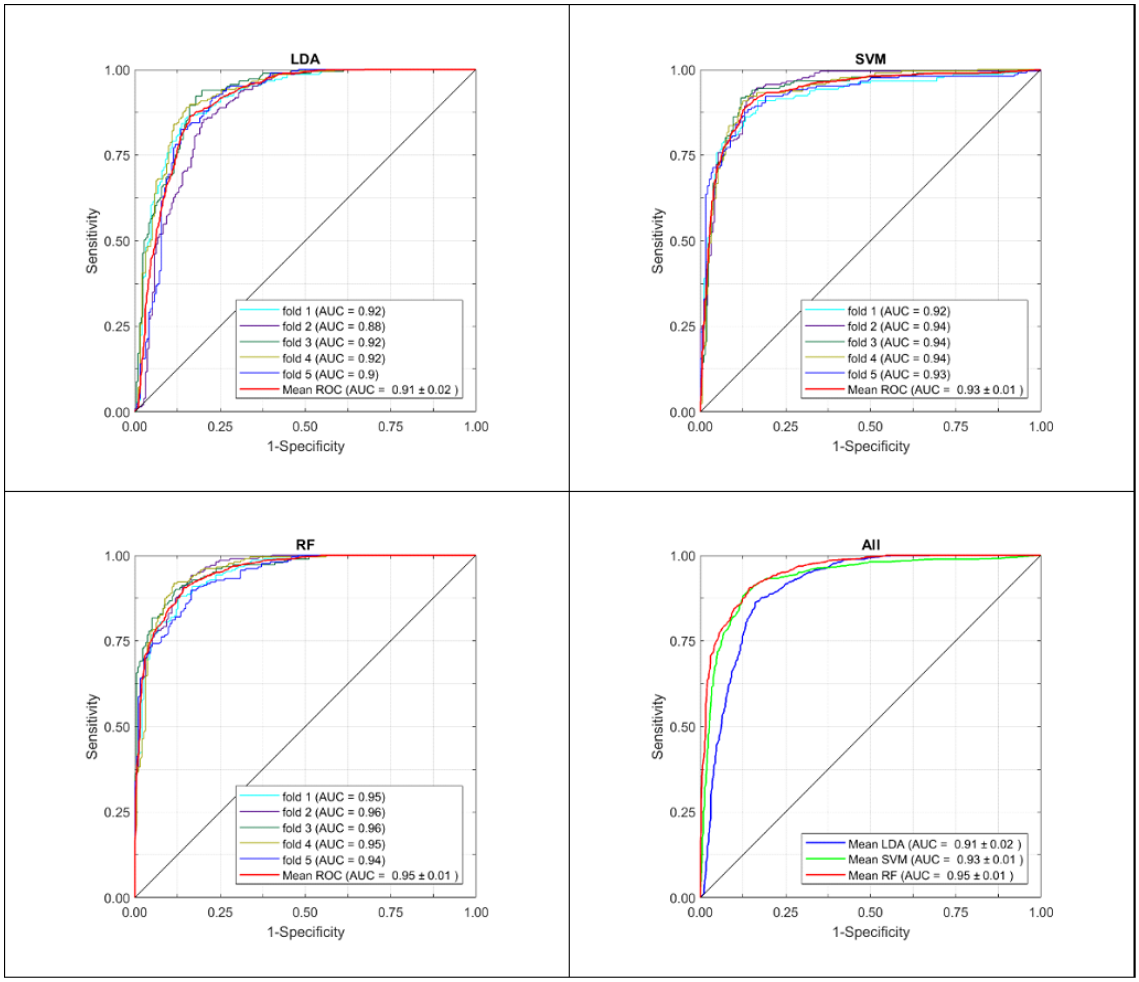


Figure 3.5 *ROC curve for classifying relevant patches versus irrelevant regions of lung adenocarinoma images using statistical image features. The RF classifier has the highest average AUC (0.95) among other applied models.*

5-fold cross validation scheme.

The ROC curve of classifying relevant and non-relevant patches from two cancer types of lung adenocarcinoma and squamous cell carcinoma is shown in Figures 3.5 and 3.6 , respectively. As results show, all tested classifiers can accurately separate relevant patches from irrelevant ones. However, for both cancer types, RF classifier has better performance in all folds in terms of AUC factor.

As described, the aim of the designed model for pre-processing step is to select 10 relevant patches from each histopathology image. In the current problem, we defined relevant patches as positive class and irrelevant patches such as tissue folds or sparse regions as negative class. Table 3.1 illustrates performance of three tested classifier for detecting relevant patches of images from lung adenocarcinoma dataset. Similarly, the performance of tested classifiers on squamous cell carcinoma dataset

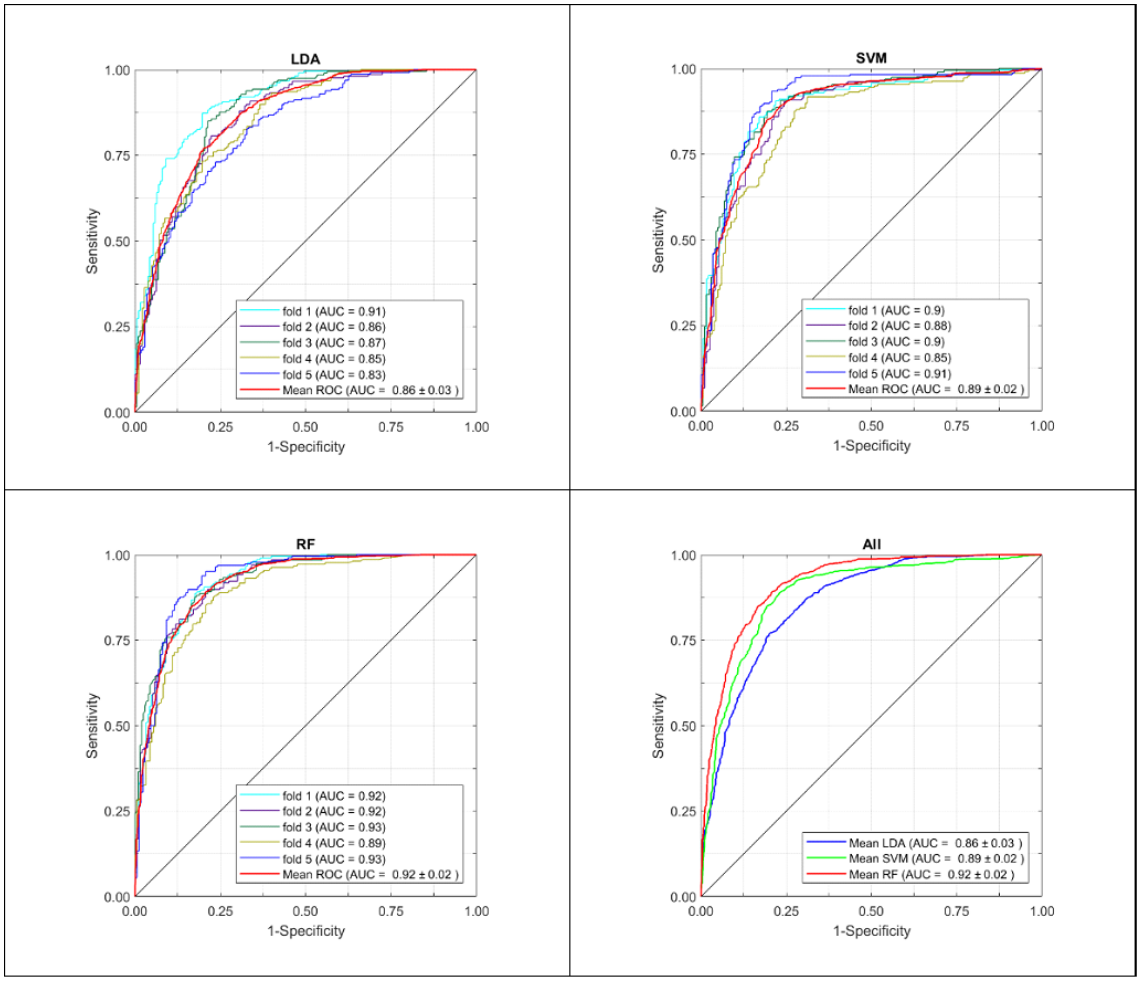


Figure 3.6 *ROC curve for classifying relevant patches versus irrelevant regions of squamous cell carcinoma images using statistical image features. The RF classifier has the highest average AUC (0.92) among other applied models.*

are compared in Table 3.2.

Sensitivity, or True Positive rate, is the portion of positive samples that are truly classified as positives. Thus, in the current problem, it indicated the number of patches that are detected as relevant to the total number of truly relevant patches. For both cancer types, the LDA classifier has higher Sensitivity values.

However, as the whole-slide images are large-sized pictures with high resolutions, there are numerous possible options within each histopathology image to be selected as a relevant patch of size 1000×1000. Therefore, we do not need to worry about some relevant patches which might be wrongly detected as irrelevant ones and be discarded by the classification system. In contrast, false positive alarms are important to our problem as they indicate the number of irrelevant patches that are wrongly detected as relevant sections and were stored in the dataset for further pro-

cessing steps. For example, selecting patches from tissue fold regions would produce wrong data for the final classification system with the task of separating cancerous tissues from healthy ones. Therefore, a classifier that produces lower false positive alarms is more suitable for the pre-processing step.

Specificity, or True Negative rate, is the portion of irrelevant patches that are correctly classified as irrelevant. As described above, we need our classifier to have high specificity to be able to perform well in detecting and avoiding irrelevant patches. As it can be seen from Tables 3.1 and 3.2, for both cancer types, RF and SVM algorithms result in higher Specificity compared to LDA classifier.

As a conclusion, RF classifier has the better performance among other tested algorithms for selecting informative regions of histopathology images of lung adenocarcinoma and squamous cell carcinoma. Therefore, for each cancer type, we trained an RF classifier using the dataset we assigned previously for preprocessing step

Finally, we used the trained RF classifiers on the whole dataset of histopathology images to detect 10 relevant patches from each whole-slide image. The selection of relevant patches was applied using a search algorithm; For each image, the code starts to detect 1000×1000 patches from random positions. For each detected patch, the trained RF classifier decides if the selected patch is a relevant area or not. If relevant, the algorithm saves the patch and add the number of selected patches by one. If irrelevant, the algorithm ignores the patch and adds the number of discarded patches by one. This process continues till either 10 relevant patches are selected, or number of discarded patches reach a maximum defined by the code to avoid spending too much time on the images mostly covered by non-relevant patches such as artifacts or tissue folds.

3.3 Feature extraction

3.3.1 CellProfiler

CellProfiler is an open-source modular software designed for analysis of biological images using advanced algorithms [45] [46]. The code has GNU public license; thus, all users have the possibility to access underlying methods and algorithms and modify them. CellProfiler accepts and supports the most common two-dimensional image formats and produces the output data vectors in the form of MATLAB or HDF5 files. CellProfiler allows to select and design a pipeline of any sequential set of individual modules. Each module is designed to apply some specified tasks

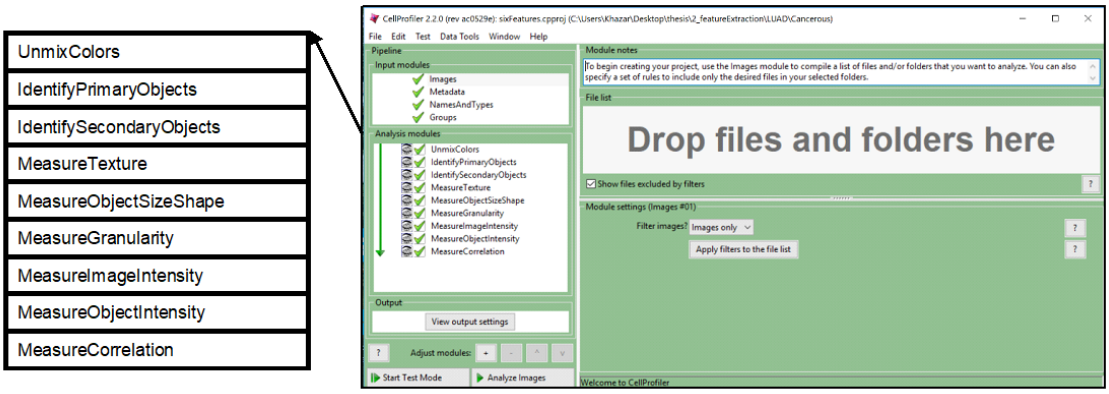


Figure 3.7 *CellProfiler pipeline used in feature extraction process. The "UnmixColors" module deconvolves Hematoxylin and Eosin channels. The second and third modules identify nuclei and cytoplasm of the cells and the next following modules measure various quantitative features from segmented images and cell objects.*

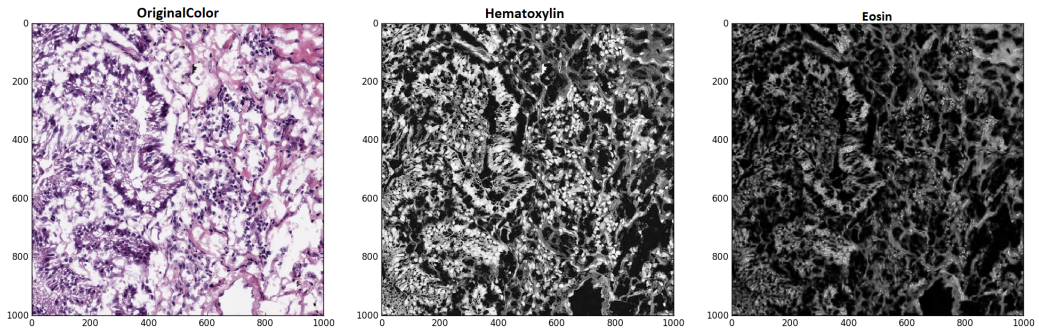


Figure 3.8 *The "UnmixColors" module splits the input HE stained image (left) to its component. The Hematoxylin channel (middle) resembles nuclie of the cells and Eosin channel (right) illustrates cytoplasm regions.*

using various algorithms. In most modules, it is possible for the user to adjust different parameters and functions to fulfill the desired goals of the defined projects. Different image processing and object processing modules are available to measure the statistics of a microscopic image. CellProfiler allows the user to measurements various attributes such as size, shape, count, and distribution and localization of the cells and other organelles as well as different texture, intensity features of the microscopic images. The task of object processing mostly includes a set of selected measurements such as size, shape, number, texture, and intensity applied on input objects. As an example, CellProfiler allows to apply different algorithms for measuring cell morphology, which is one of the main important factors in quantifying

physiology of the cell and diagnosis of some certain diseases. Moreover, CellProfiler applies necessary adjustments such as illumination correction or image alignment on the primary images before performing further analysis and measurements. Figure 3.7 shows the CellProfiler pipeline used at this work.

3.3.2 Segmentation of tissue objects

The first important step in analyzing any microscopic image is detection and identification of objects of interest. There are certain object identification modules in CellProfiler for detecting cells and other organelles. As an advantage, the CellProfiler modules use advanced algorithms to detect overlapping cells and object.

The segmentation modules of the CellProfiler allow the user to adjust different parameters for identifying primary and secondary objects. These parameters include estimated diameter of the objects, threshold strategy, and the method to distinguish clumped objects.

In general, defining primary objects like nuclei yields to recognition of secondary surrounding objects such as cells. However, most segmentation methods such as watershed algorithm fail to detect clumped objects with a microscopic image. Therefore, a three-step strategy is applied in CellProfiler to identify clumped objects. Firstly, clumped objects are segmented using a segmentation method such as watershed. Secondly, the algorithm tries to identify the dividing lines between the recognized objects. Finally, some objects are discarded or merged again using previously obtained measurements.

As mentioned before, the histopathology images used in this research are stained using HE method in which nuclei of the cells appear in blue color and cytoplasm appear in red or pink color. Figure 3.8 shows an image patch in original color as well as the results obtained after applying Unmixcolor module in CellProfiler. As it can be seen from the Figure, in Hematoxylin channel, the nuclei of the cells and the background region appear in white and black, respectively, and the area occupied by cytoplasm has shades of gray.

We used Watershed algorithm for identifying the secondary object since it is the most widely used method for segmentation of the cell objects. As it can be seen from Figure 3.7, the first 3 applied modules in the designed CellProfiler pipeline are segmentation modules used to separate HE color components and identify cells and nuclei of the cells.

- "UnmixColors"

  This module segments the Hematoxylin and Eosin staining channles producing two grayscale images. Hematoxylin and Eosin channels represent the nuclei and cytoplasm of the cells, respectively. Figure.1 shows an example of the outputs of UnmixColors module for an HE stained histopathology image of the lung tissue.

- "IdentifyPrimaryObjects"

  The output of the Hematoxylin channel is fed to this module to segment and identify the nuclei of the cells.

- "IdentifySecondaryObjects"

  This module uses the Eosin image as well as nuclei objects as input to define the cell regions.

3.3.3 Extracting biologically relevant features

After segmenting tissue objects, several features were derived using 6 feature extraction modules. These modules apply various evaluation methods to measure texture, intensity, and correlation attributes of the images, as well as, texture, intensity, morphology and granularity of the segmented objects.

- "MeasureTexture"

  Haralick features and Gabor features of the images as well as primary and secondary objects can be measured using MeasureTexture module. This module measures variations of pixel intensities within any given image and derives object texture using grayscale images. There are options for adjusting angle and pixel scale of the measured features.

- "MeasureObjectSizeShape"

  This module measures area and different shape and morphological features of cells and nuclei of the cells. These features include Zernike shape features, area, perimeter, and solidity of the objects, Euler number, eccentricity and orientation of the ellipse, etc.

- "MeasureGranularity"

  This module calculates granularity of any input image by fitting a series of structure kernels of increasing size to find the size of objects presented in

the image. In this work, we used this module to measure granularity of the Hematoxylin and Eosin channels. Sub-sampling factor was adjusted to default value of 0.25 since the images are provided in high resolution. To avoid background variation to affect the granularity measurement, a sub-sampling factor is applied to normalize image volume at certain granular size by total image volume.

- "MeasureImageIntensity"

  This module was used to measure intensity features within Hematoxylin and Eosin channels. These features include total intensity, mean intensity, standard deviation, total area, etc.

- "MeasureObjectIntensity"

  This module accepts an image and a corresponding object as input and extracts intensity features for each of the segmented objects based on their correlated image.

- "MeasureCorrelation"

  We used this module to measure correlation between Eosin and Hematoxylin channels across the entire image.

3.4 Classification

The last step in the classification pipeline includes applying machine learning algorithms on the obtained data vectors to decide whether a given whole-slide microscopy image is benign or malignant.

The classifiers used for this task include Linear LDA, Ada boost, and RF. To apply several classification methods on the obtained data vectors, we used scikit-learn, NumPy and SciPy libraries in Python. In this section, we firstly describe the methodology used at classification step. Next, the efficiency of the applied pipeline is assessed using various performance measurement metrics.

3.4.1 Classification methods

We used Python 3.6 as the programming environment for applying classification algorithms on the data vectors obtained in previous step. All necessary packages were downloaded and added using anaconda package manager tool.

Table 3.3 *Distribution of samples in the final data vectors*

Image type	Number of features	Number of patches	Number of groups	Number of malignant samples	Number of normal samples
lung adenocarcinoma	1400	9876	1021	7927	1949
squamous cell carcinoma	1400	9839	1001	6800	3039

Different operations were applied on the available data set using scikit-learn and NumPy packages in python. Scikit-learn is a free machine learning library written in python which provides the user with the various classification, regression, and clustering algorithms including linear and non-linear models. NumPy is a python library that supports multi-dimensional arrays of large scaled data and a large set of mathematical operators and functions. Additionaly, SciPy library is used to load feature vectors from MATLAB data files.

After applying pre-processing method, we used 1021 and 1001 whole-slide histopathology images from each of lung adenocarcinoma and squamous cell carcinoma cancer types respectively. Using CellProfiler, we extracted 1400 features from each image patch. Thus, the feature vector had size of 1400 by 9876 for lung adenocarcinoma data and 1400 by 9839 for squamous cell carcinoma data.

We labeled Cancerous images as class 0 and the images obtained from adjacent normal tissues as class 1. Moreover, the image patches cropped from each histopathology image were collected in one group. Thus, the number of groups demonstrates the number of employed histopathology whole-slide images.

A 5-fold cross validation method was applied in a group-wise manner to separate data to two sets of training and testing. The classifiers were trained using training data and their performances were evaluated using testing data set. Moreover, the image patches derived from each histopathology WSI were labeled with the same group number. Storing the group labels of images prevents the patches cropped from the same histopathology image to appear in training and testing sets simultaneously.

As described in chapter 2, in a cross-validation assessment technique, the data is randomly partitioned to k almost equal subsets and in each turn, the classifier is trained using k-1 subsets of the data and it performance is tested using the remaining one subset. Thus, with a 5-fold cross validation, 1 fifth of the whole data (0.2) is considered for testing and the algorithm is trained using 4/5 of the data set (0.8).

From skit-learn package, model-selection and metrics libraries were used to evaluate the performance of the applied classifiers on testing data. Various assessment techniques were applied to measure the performance of our classifiers.

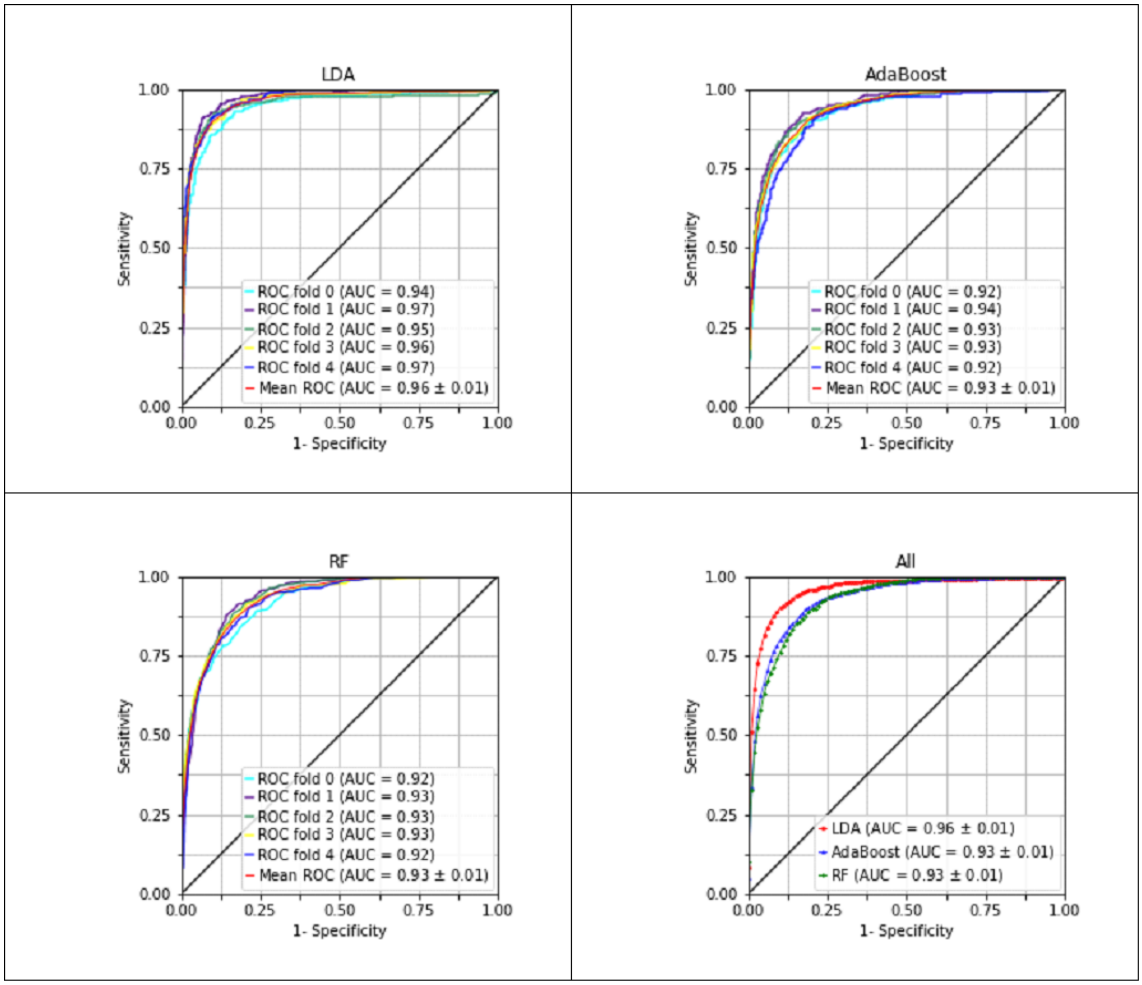


Figure 3.9 *ROC curve of classifying malignant and healthy tissues in lung adenocarcinoma patients.*

Table 3.4 *Quantitave evaluation of different models for classifying malignant cells versus adjacent normal tissues in lung adenocarcinoma patients*

Method	Accuracy	Precision	Recall	F1 measure	AUC
LDA	**0.92±0.01**	**0.82±0.03**	**0.78±0.01**	**0.80±0.02**	**0.96**±0.01
AdaBoost	0.89±0.01	0.76±0.03	0.67±0.04	0.71±0.03	0.93±0.01
RF	0.88±0.01	0.81±0.02	0.53±0.03	0.64±0.02	0.93±0.01

3.4.2 Results and discussions

Table 3.4 and 3.5 compare different performance scores of the applied classification models on lung adenocarcinoma and squamous cell carcinoma data, respectively. The tables illustrated the average and standard deviation of different performance metrics of the 5 testing folds in the applied 5-fold cross-validation method. The measured scores include Accuracy, Precision, Recall, F1- score, and AUC. Moreover, the ROC curve obtained for classifying lung adenocarcinoma and squamous cell

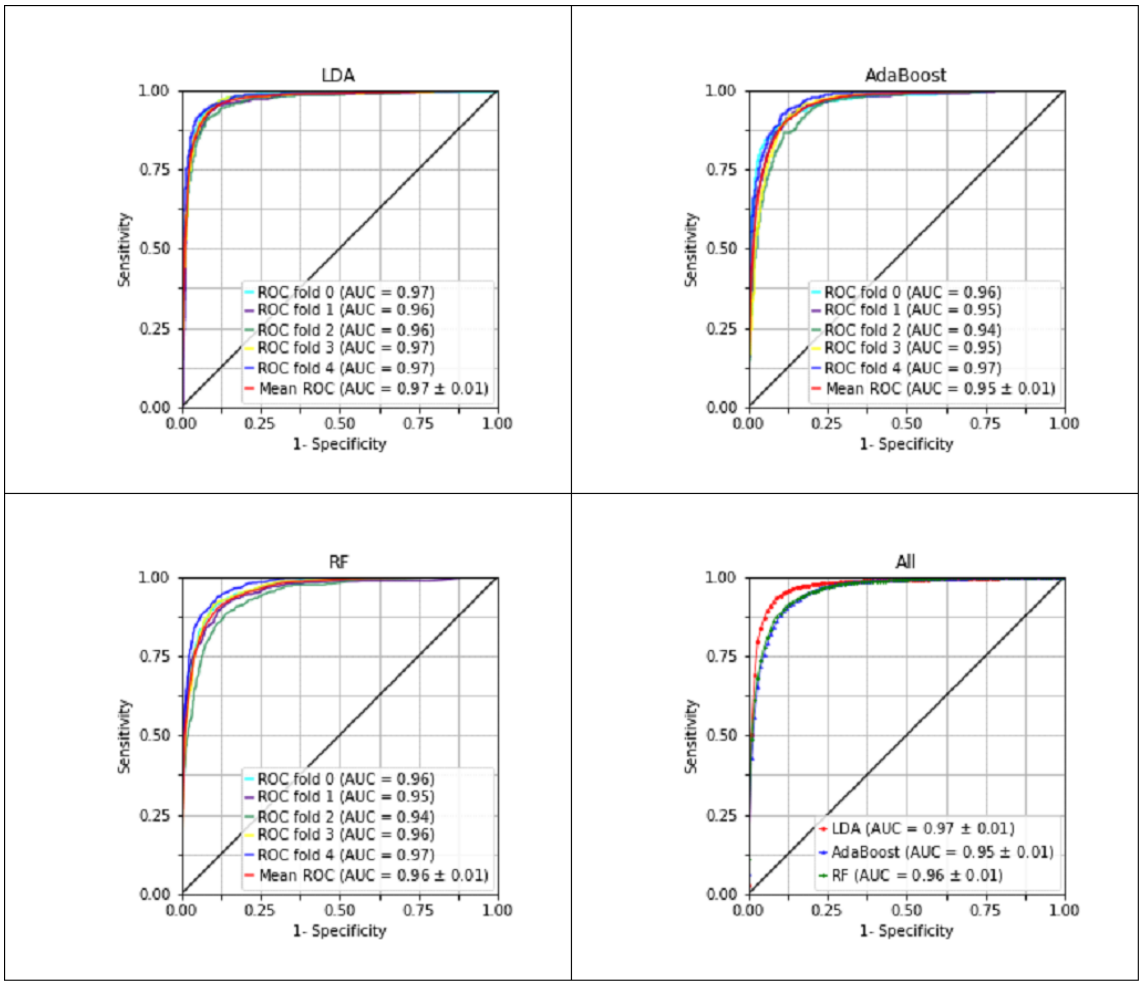


Figure 3.10 *ROC curve of classifying malignant and healthy tissues in squamous cell carcinoma patients.*

Table 3.5 *Quantitave evaluation of different models for classifying malignant cells versus adjacent normal tissues in lung squamous cell carcinoma patients*

Method	Accuracy	Precision	Recall	F1 measure	AUC
LDA	**0.92±0.01**	**0.88±0.03**	**0.87±0.02**	**0.88±0.01**	**0.97±0.01**
AdaBoost	0.90±0.01	0.85±0.04	0.82±0.01	0.83±0.02	0.95±0.01
RF	0.90±0.01	0.87±0.03	0.79±0.02	0.83±0.02	0.96±0.01

carcinoma images, are presented in Figure 3.9 and 3.10, respectively.

Considering all the measured scores, the LDA classifier has the best performance among other tested models in both cancer types, showing that the attributes we calculated previously from image patches of histopathology slides are dividable linearly. As results show, for classifying between malignant tissues and healthy adjacent cells in lung adenocarcinoma and squamous cell carcinoma patients, we achieved AUC of 0.96 and 0.97, respectively.

4. CONCLUSIONS

In this thesis, we proposed a solution for improving the efficiency of automated systems for discriminating between malignant and healthy patients. The performance of the proposed methodology was assessed on haematoxylin and eosin stained histopathology whole slide images (WSIs) of lung adenocarcinoma and squamous cell carcinoma patients obtained from The Cancer Genome Atlas (TCGA) dataset.

A fully automated pipeline was designed for classifying images containing malignant cells from those obtained from adjacent normal tissues. In preprocessing step, an RF algorithm used to detect 10 patches from clinically relevant sections of WSIs based on the statistical measurement of each patch. Next, a set of biologically significant features were extracted from selected patches using CellProfiler software. In the final step, various ML models were applied to distinguish between cancerous and non-cancerous tissue types.

Applying ML algorithms at pre-processing step for selecting clinically relevant sections of histopathology images helps to accelerate processing of huge amount of data exists in WSIs. Furthermore, it increases the accuracy of automated cancer detection systems significantly via discarding diagnostically irrelevant regions such as tissue folds or sparse areas.

As results show, the set of extracted features accurately distinguish between benign and malignant cells of lung adenocarcinoma and squamous cell carcinoma patients. However, the discrimination between the two cancer types should still be studies. While the similar pipeline might be used for this task, the number of measured features required to be increase significantly due to the complexity of the tissue structures and similarity between cancerous patterns presented in histopathology images of these two malignancy types.

In this work, the effectiveness of the proposed method was tested only on two types of lung cancer patients. However, the cancerous cells show various visual patterns while attacking different body organs and tissue types. Therefore, the effectiveness of the proposed methodology should be studies for a wider range of malignancies

where histopathology image analysis are applied for cancer diagnosis such as various breast cancer types.

In addition, applying CellProfiler software for extracting various clinically significant image features is computationally expensive and demands hours of heavy workload. As an alternative method, one might utilize deep learning (DL) methods such as convolutional neural networks (CNNs) as a feature extraction tool. As it is possible to run CNN models on a graphics processing unit (GPU), it might result in faster implementation on the feature-extraction and classification pipelines. Therefore, the cropped patches obtained in pre-processing step can be directly fed to a CNN algorithm which is designed to recognize between malignant and benign tissue images. Moreover, using DL models allows adding the meta-data stored in patients files such as age, gender, and race of the patients as an input parameter to the feature extraction pipeline. Therefore, it enables studying the effect of these attributes in diagnosis of cancer or differentiating between various cancer types.

BIBLIOGRAPHY

[1] He, Lei, et al. "Histology image analysis for carcinoma detection and grading." Computer methods and programs in biomedicine 107.3 (2012): 538-556.

[2] Fuchs, Thomas J., and Joachim M. Buhmann. "Computational pathology: Challenges and promises for tissue analysis." Computerized Medical Imaging and Graphics 35.7-8 (2011): 515-530

[3] Peikari, Mohammad, et al. "Triaging diagnostically relevant regions from pathology whole slides of breast cancer: A texture based approach." IEEE transactions on medical imaging 35.1 (2016): 307-315.

[4] Bahlmann, Claus, et al. "Automated detection of diagnostically relevant regions in HE stained digital pathology slides." Medical Imaging 2012: Computer-Aided Diagnosis. Vol. 8315. International Society for Optics and Photonics, 2012.

[5] Palokangas, Sakari, Jyrki Selinummi, and Olli Yli-Harja. "Segmentation of folds in tissue section images." Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE. IEEE, 2007.

[6] Bautista, Pinky A., and Yukako Yagi. "Detection of tissue folds in whole slide images." Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE. IEEE, 2009.

[7] Ojansivu, Ville, et al. "Automated classification of breast cancer morphology in histopathological images." Diagnostic Pathology. Vol. 8. No. 1. BioMed Central, 2013.

[8] Kumar, Rajesh, Rajeev Srivastava, and Subodh Srivastava. "Detection and classification of cancer from microscopic biopsy images using clinically significant and biologically interpretable features." Journal of medical engineering 2015 (2015).

[9] Yu, Kun-Hsing, et al. "Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features." Nature communications 7 (2016): 12474.

[10] Carpenter, Anne E., et al. "CellProfiler: image analysis software for identifying and quantifying cell phenotypes." Genome biology 7.10 (2006): R100.

[11] Mitchell, Tom Michael. The discipline of machine learning. Vol. 3. Carnegie Mellon University, School of Computer Science, Machine Learning Department, 2006

[12] SEER Training Modules, Cancer as a Disease. U. S. National Institutes of Health, National Cancer Institute. 26, Nov. 2017, Available: https://training.seer.cancer.gov/disease/

[13] Baba AI, Ctoi C. Comparative Oncology. Bucharest: The Publishing House of the Romanian Academy; 2007. Available from: https://www.ncbi.nlm.nih.gov/books/NBK9557/

[14] National Cancer Institute. Available: `https://visualsonline.cancer.gov/details.cfm?imageid=2512`. (2017). [image].

[15] National Cancer Institute. Available: `https://visualsonline.cancer.gov/details.cfm?imageid=2119`. (2017). [image].

[16] Sloan, Kevin E., et al. "CD155/PVR plays a key role in cell motility during tumor cell invasion and migration." BMC cancer 4.1 (2004): 73

[17] Haralick, Robert M., and Karthikeyan Shanmugam. "Textural features for image classification." IEEE Transactions on systems, man, and cybernetics 6 (1973): 610-621.

[18] Partio, Mari, et al. "Rock texture retrieval using gray level co-occurrence matrix." Proc. of 5th Nordic Signal Processing Symposium. Vol. 75. 2002.

[19] Anuradha, K. "Statistical Feature extraction to classify oral cancers." Journal of Global Research in Computer Science 4.2 (2013): 8-12.

[20] Ahonen, Timo, Abdenour Hadid, and Matti Pietikainen. "Face description with local binary patterns: Application to face recognition." IEEE transactions on pattern analysis and machine intelligence 28.12 (2006): 2037-2041.

[21] Pietikinen, Matti, et al. "Local binary patterns for still images." Computer vision using local binary patterns. Springer London, 2011. 13-47.

[22] Ojansivu, Ville, and Janne Heikkilä. "Blur insensitive texture classification using local phase quantization." International conference on image and signal processing. Springer Berlin Heidelberg, 2008.

[23] Gonzalez, Rafael C., and Richard E. Woods. "Image segmentation. " Digital image processing. Prentice Hall, 2008. 689-795

[24] Jain, Anil K., and Sushil Bhattacharjee. "Text segmentation using Gabor filters for automatic document processing." Machine Vision and Applications 5.3 (1992): 169-184

[25] Grigorescu, Simona E., Nicolai Petkov, and Peter Kruizinga. "Comparison of texture features based on Gabor filters." IEEE Transactions on Image processing 11.10 (2002): 1160-1167.

[26] Chaudhuri, Debasis. "Global Contour and Region Based Shape Analysis and Similarity Measures." Defence Science Journal 63.1 (2013): 74.

[27] Yang, Mingqiang, Kidiyo Kpalma, and Joseph Ronsin. "A survey of shape feature extraction techniques." (2008): 43-90.

[28] Rangamani, Padmini, et al. "Decoding information in cell shape." Cell 154.6 (2013): 1356-1369.

[29] Fisher, Ronald A. "The use of multiple measurements in taxonomic problems." Annals of human genetics 7.2 (1936): 179-188.

[30] Welling, Max. "Fisher linear discriminant analysis." Department of Computer Science, University of Toronto 3.1 (2005).

[31] Hastie, Trevor, Jerome Friedman, and Robert Tibshirani. "Linear methods for classification." The Elements of Statistical Learning. Springer New York, 2009. 101-135.

[32] Webb, Andrew R., and Keith D. Copsey. "Rule and Decision Tree Induction." Statistical Pattern Recognition, Third Edition: 322-360.

[33] Hastie, Trevor, Jerome Friedman, and Robert Tibshirani. "Random Forest." The Elements of Statistical Learning. Springer New York, 2009. 587-603.

[34] Webb, Andrew R., and Keith D. Copsey. "Ensemble methods." Statistical Pattern Recognition, Third Edition: 361-403.

[35] Webb, Andrew R., and Keith D. Copsey. "Performance Assessment." Statistical Pattern Recognition, Third Edition: 404-432.

[36] Hastie, Trevor, Jerome Friedman, and Robert Tibshirani. "Model Assessment and Selection ." The Elements of Statistical Learning. Springer New York, 2009. 219-257.

[37] Bekkar, Mohamed, Hassiba Kheliouane Djemaa, and Taklit Akrouf Alitouche. "Evaluation measures for models assessment over imbalanced data sets." Iournal Of Information Engineering and Applications 3.10 (2013).

[38] Cancer Genome Atlas Research Network. "Comprehensive molecular profiling of lung adenocarcinoma." Nature 511.7511 (2014): 543-550.

[39] Cancer Genome Atlas Research Network. "Comprehensive genomic characterization of squamous cell lung cancers." Nature 489.7417 (2012): 519-525.

[40] Available:https://portal.gdc.cancer.gov/legacy-archive/search/f

[41] Grossman, Robert L., Heath, Allison P., Ferretti, Vincent, Varmus, Harold E., Lowy, Douglas R., Kibbe, Warren A., Staudt, Louis M. (2016) Toward a Shared Vision for Cancer Genomic Data. New England Journal of Medicine375:12, 1109-1112

[42] Linkert, Melissa, et al. "Metadata matters: access to image data in the real world." The Journal of cell biology 189.5 (2010): 777-782.

[43] Available: https://www.openmicroscopy.org/bio-formats/

[44] Available: https://downloads.openmicroscopy.org/bio-formats/5.0.0/artifacts/Bio-Formats-5.0.0.pdf

[45] Jones, Thouis R., et al. "CellProfiler Analyst: data exploration and analysis software for complex image-based screens." BMC bioinformatics 9.1 (2008): 482.

[46] Lamprecht, Michael R., David M. Sabatini, and Anne E. Carpenter. "CellProfiler: free, versatile software for automated biological image analysis." Biotechniques 42.1 (2007): 71.